2019

Nonparametric curve estimation and applications

Yu Liu
Iowa State University

Nonparametric curve estimation and applications

by

Yu Liu

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Computer Science

Program of Study Committee:
Kris De Brabanter, Major Professor
Daniel J. Nordman
Daniel S. Nettleton
Jin Tian
Samik Basu

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2019

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to express my gratitude to my colleagues, family and friends, without whose support I would not have complete this work.

First of all, I would like to thank my parents, who give me love and financial support.

Second, I would like to thank my major Professor Dr. De Brabanter for giving me the opportunity to work with him after the retirement of my first advisor. I am very appreciative for his guidance in my research, continuous financial support, patience in answering my questions and relieving my pressure during my Ph.D. He inspired my enthusiasm for my research and gave me the freedom to pursue a Master degree in Statistics.

Third, I would like to thank Dr. Weiss who supported me during my two years of my program, to Dr. Attinger who guided me in the bloodstain pattern analysis project, to Dr. Nettleton who mentored me in the maize yield project. I also would like to thank Dr. De Brabanter, Dr. Nettleton, Dr. Nordman, Dr. Tian and Dr. Basu for their insights and guidance in writing this manuscript.

Fourth, I would like to thank my boyfriend Yetian Chen for his accompany on this journey. I would like to my friends and colleagues, Yu Jie, Qiuyan Liao, Chaohui Yuan, Chuan Jiang, WangYujue Hong, Xiang Huang, Li Lin and all my friends in Ames and China for making me feel happy and energetic.

ABSTRACT

Nonparametric methods are popular in data analysis since it requires few assumptions about the underlying populations that the data are coming from, e.g. nonparametric regression and nonparametric density estimation. Nonparametric regression drew a lot of attention since mid sixties and it is well studied because it provides the possibility to uncover the nonlinear relationship between a dependent variable and one or more independent variables without imposing the assumption on the shape of the mean function. We propose simultaneous confidence intervals for the estimated regression curve using nonparametric methods. We also propose a nonparametric derivative estimation method and apply the nonparametric method in bloodstain pattern analysis.

First, we study the convergence rate of simultaneous bias corrected confidence intervals for a smooth curve using local polynomial regression which is a well-known nonparametric regression technique in the area of statistics. We extend the idea of volume-of-tube to construct the simultaneous confidence intervals for this biased estimator. We empirically show that the proposed simultaneous confidence intervals attain, at least approximately, nominal coverage.

Second we propose nonparametric first and second order derivative estimators without having to estimate the regression function. The estimator is based on a variance-reducing linear combination of symmetric difference quotients. We establish the asymptotic properties of the proposed derivative estimators and propose fast tuning methods to select parameters. We compare the proposed estimators with popular estimators for derivative estimation such as local polynomial regression and smoothing splines.

Last we apply the nonparametric methods to construct the features in classification of bloodstain patterns. Bloodstain pattern analysis (BPA) plays an important role in foren-

sics towards crime scene analysis. We propose an automated framework to classify the bloodstain spatters caused by either gunshot or blunt impact, based on machine learning methods. We analyze 94 blood spatters which are being disseminated as free public datasets for research purposes and construct features using nonparametric methods. The study also shows how the distance between the target surface collecting the stains and the blood source influences the bloodstain pattern. Finally we obtain the accuracy of the proposed classification model for different distances ranges.

CHAPTER 1. INTRODUCTION

1.1 General background of nonparametric methods

Nonparametric methods require few assumptions regarding the underlying distribution which the data are obtained (Hollander and Wolfe, 1999). The old nonparametric methods can be traced back to 1710, when John Arbuthnot used the sign test to analyze the birth records in London (Conover and Conover, 1980; Sprent and Smeeton, 2000). Other early famous work include Pearson's χ^2 test proposed by Karl Pearson to test goodness of fit and homogeneity (Pearson, 1900, 1911). Nonparametric methods have grown dramatically since mid 1930s due to many advantages they enjoy, e.g. relative mild assumptions, relative simplicity and relative insensitive to outlying observations. Extensive research was conduct with regards to the nonparametric statistical hypothesis test. Friedman (1937) used ranks of data to analyze the variance without normality assumption. Wilcoxon (1945) proposed the Wilcoxon signed-rank test to compare the matched samples for testing the significance of the difference of the mean between them. In the same paper, Wilcoxon also introduced the Wilcoxon rank-sum test for two independent samples with equal sample size and Mann and Whitney (1947) gave the general result when sample size are not equal and discussed the distribution of the statistics. There are also two famous nonparametric tests for goodness of fit and comparing two samples: Kolmogorov-Smirnov test which was initially proposed in Kolmogoroff (1941); Smirnov (1939) and Cramér-Von Mises test (Anderson, 1962; Cramér, 1928; Darling, 1957).

Important achievements of nonparametric methods are also made by introducing new resampling techniques, e.g. jackknife and bootstrap. Quenouille introduced a resampling

technique jackknife to reduce the bias of a serial correlation estimator in a nonparametric way (Quenouille, 1949) and he generalized this idea later (Quenouille, 1956). Tukey proposed the robust interval estimation based on jackknife technique (Tukey, 1958; Miller, 1974). Efron et al. proposed the nonparametric bootstrap which provides the straightforward nonparametric way to estimate the properties of an estimator, e.g. standard errors and confidence intervals, in the setting that it is difficult to do that in parametric approach (Efron, 1992; Efron and Tibshirani, 1994). In addition, they showed that the jackknife can be thought of as a linear expansion method for approximating the bootstrap.

Another important research area is Bayesian nonparametric statistics, in which nonparametric means model with infinite-dimensional parameters. Cox (1972) introduced a semi parametric model, which is commonly used in survival analysis. Ferguson (1973) proposed the famous Dirichlet process and later Polya Trees prior was considered by (Ferguson et al., 1974; Blackwell et al., 1973). Other work include the Markov Chain Monte Carlo (MCMC) methods (Gilks et al., 1995) e.g. Metropolis-Hastings (Metropolis et al., 1953; Hastings, 1970) and Gibbs sampling (Gelfand and Smith, 1990; Casella and George, 1992; Tierney, 1994).

1.2 Literature Review of Nonparametric Curve Estimation

In this dissertation, we focus more on nonparametric curve estimation, specifically kernel based estimators which have been popular in the field of statistics and machine learning. The real breakthrough of kernel based nonparametric curve estimation was made in 1950s and 1960s. Parzen (1962) and Rosenblatt (1956) proposed the initial work on kernel density estimation and its asymptotic distribution was studied by Bickel and Rosenblatt (1973). A various bandwidth selection methods for kernel density estimation, e.g. plug-in method, normal-reference method, cross-validation method and so on, were proposed (Chiu, 1991; Hall et al., 1991; Woodroofe, 1970; Jones et al., 1996; Bowman, 1984). A comprehensive topics of density estimation can be found in Silverman (2018) and Scott (2015).

Another important topic of nonparametric curve estimation is the regression curve estimation, which aims to find the relationship between the dependent variable and one or more independent variables. The regression estimation has a long history. The famous least squares method for linear regression was proposed by Legendre in 1805 and Gauss in 1823 and 1828 (Legendre, 1805). After a long time, parametric regression is popular due to its simplicity in computation until at some point, it turns out that the assumption of certain parametric form, e.g. linear or quadratic, is not suitable to explain the data. Thus, nonparametric regression methods which provide more flexible models are used for empirical analysis in 19th century. The fundamental Nadaraya-Watson regression estimator independently proposed by Nadaraya and Watson (Nadaraya, 1964; Watson, 1964) attracted attention in nonparametric regression methods. Researchers realized that the nonparametric regression estimators provides more flexibility in data analysis and the development of computing hardware supports the complex computation in nonparametric regression estimators. Plenty of distinguished works were emerged, e.g. Gasser-Müller estimator (Gasser and Müller, 1979), wavelet-based method (Donoho, 1995; Donoho and Johnstone, 1994; Donoho et al., 1998; Donoho and Johnstone, 1995), spline method (Eubank, 1988; Wahba, 1990; Green and Silverman, 1993) and local polynomial regression (Stone, 1977, 1980, 1982; Cleveland, 1979; Fan, 1992, 1993). Similar to the density estimation, considerable work has also been devoted to the bandwidth selection problem of regression estimation. Some parameter selection techniques can also be applicable to both density estimation setting and regression estimation setting. Most commonly used approaches include the cross-validation and plugin-in methods (Hall and Johnstone, 1992; Härdle, 1990; Fan and Gijbels, 1996; Ruppert et al., 1995). Though tremendous work has been done on bandwidth selection, there is no general good method suitable for all cases, e.g. Fan mentioned "Least squares cross-validation suffers from a lot of sample variation and searching the minimum of the CV functions can cause problem" (Fan and Gijbels, 1996). Thus, in the Chapter 3 of this dis-

sertation, we propose a rule of thumb tuning parameter method along with the estimators, instead of using cross validation.

Asymptotic properties of the regression estimator and data-driven bandwidth selection under the independent assumptions among samples break down in the presence of correlated errors. In order to establish the asymptotic normality of local polynomial regression estimators, additional assumptions are necessary (Francisco-Fernández and Vilar-Fernández, 2001). In addition, bandwidth selection methods assuming independent errors (Fan and Gijbels, 1996) perform poorly if the independency assumption is violated. In this dissertation, we applied the bandwidth selection method in De Brabanter et al. (2018) to deal with correlated errors.

1.3 Organization and Contributions of The Thesis

The main body of this dissertation is organized into three chapters, one for each individual topic. Chapter 2 and Chapter 3 establish some theoretical results with regards to the nonparametric curve estimation. Chapter 4 applies nonparametric curve estimation techniques to conduct blood pattern analysis.

In Chapter 2, we construct the bias-corrected uniform (simultaneous) intervals for local polynomial regression curve, based on the volume-of-tube formula. We study the asymptotic order of the width of this bias-corrected uniform intervals for p odd and p even in the interior. The asymptotic order remains the same in the boundary region for p odd, while for p even the asymptotic order in the boundary region is slower than in the interior. We compare the coverage probabilities and median areas of the proposed uniform confidence intervals with other three types of uniform confidence intervals. It turns out the proposed uniform confidence intervals attain roughly the nominal converge. Finally, we show the empirical performance of the proposed uniform confidence intervals for independent non-Gaussian errors and apply the proposed intervals on two real life data.

In Chapter 3, we propose the first and second order empirical (noisy) derivative estimator for the random design, based on the difference quotient. We study the asymptotic properties for the empirical derivative estimator and propose an efficient tuning parameter method. Later, we extend the work in De Brabanter et al. (2018) to establish asymptotic bias and variance decompositions and discuss the asymptotic properties of the smoothed derivative estimators. We compare the performance of the proposed derivative estimator with two popular nonparametric methods: local polynomial regression and smoothing splines.

Bloodstain spatters are a collection of blood stains and the analysis of bloodstain spatter has been used in US criminal courts. In Chapter 3, we propose an automatic quantitative and objective framework to classify the blunt impact and backward gunshot spatters by combining digital image analysis and machine learning methods. It also turns out that the distance between the blood source to the target board influences the performance of the proposed framework. During the study, we construct important features based on domain knowledge and statistical methods. Based on our data set, the proposed model achieves 98.81% accuracy in classifying spatters at distances within 30 cm, 93.20% accuracy at distances within 60 cm and 85.96% accuracy at distances within 120 cm.

1.4 References

Anderson, T. W. (1962). On the distribution of the two-sample cramer-von mises criterion. *The Annals of Mathematical Statistics*, pages 1148–1159.

Bickel, P. J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *The Annals of Statistics*, pages 1071–1095.

Blackwell, D., MacQueen, J. B., et al. (1973). Ferguson distributions via pólya urn schemes. *The annals of statistics*, 1(2):353–355.

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71(2):353–360.

Casella, G. and George, E. I. (1992). Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174.

Chiu, S.-T. (1991). Bandwidth selection for kernel density estimation. *The Annals of Statistics*, pages 1883–1905.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836.

Conover, W. J. and Conover, W. J. (1980). Practical nonparametric statistics.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Cramér, H. (1928). On the composition of elementary errors: First paper: Mathematical deductions. *Scandinavian Actuarial Journal*, 1928(1):13–74.

Darling, D. A. (1957). The kolmogorov-smirnov, cramer-von mises tests. *The Annals of Mathematical Statistics*, 28(4):823–838.

De Brabanter, K., Cao, F., Gijbels, I., and Opsomer, J. (2018). Local polynomial regression with correlated errors in random design and unknown correlation structure. *Biometrika*, 105(3):681–690.

Donoho, D. L. (1995). Nonlinear solution of linear inverse problems by wavelet–vaguelette decomposition. *Applied and computational harmonic analysis*, 2(2):101–126.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the american statistical association*, 90(432):1200–1224.

Donoho, D. L., Johnstone, I. M., et al. (1998). Minimax estimation via wavelet shrinkage. *The annals of Statistics*, 26(3):879–921.

Donoho, D. L. and Johnstone, J. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *biometrika*, 81(3):425–455.

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Eubank, R. L. (1988). *Spline smoothing and nonparametric regression*. Number 04; QA278. 2, E8.

Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American statistical Association*, 87(420):998–1004.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics*, pages 196–216.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, volume 66. CRC Press.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.

Ferguson, T. S. et al. (1974). Prior distributions on spaces of probability measures. *The annals of statistics*, 2(4):615–629.

Francisco-Fernández, M. and Vilar-Fernández, J. M. (2001). Local polynomial regression estimation with correlated errors. *Communications in Statistics-Theory and Methods*, 30(7):1271–1293.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200):675–701.

Gasser, T. and Müller, H.-G. (1979). Kernel estimation of regression functions. In *Smoothing techniques for curve estimation*, pages 23–68. Springer.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. (1995). *Markov chain Monte Carlo in practice*. Chapman and Hall/CRC.

Green, P. J. and Silverman, B. W. (1993). *Nonparametric regression and generalized linear models: a roughness penalty approach*. CRC Press.

Hall, P. and Johnstone, I. (1992). Empirical functionals and efficient smoothing parameter selection. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 475–530.

Hall, P., Sheather, S. J., Jones, M., and Marron, J. S. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika*, 78(2):263–269.

Härdle, W. (1990). *Applied nonparametric regression*. Cambridge university press.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications.

Hollander, M. and Wolfe, D. A. (1999). Nonparametric statistical methods.

Jones, M. C., Marron, J. S., and Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American statistical association*, 91(433):401–407.

Kolmogoroff, A. (1941). Confidence limits for an unknown distribution function. *The Annals of Mathematical Statistics*, 12(4):461–463.

Legendre, A. M. (1805). *Nouvelles méthodes pour la détermination des orbites des comètes*. F. Didot.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, pages 50–60.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Miller, R. G. (1974). The jackknife-a review. *Biometrika*, 61(1):1–15.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076.

Pearson, K. (1900). X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175.

Pearson, K. (1911). On the probability that two independent distributions of frequency are really samples from the same population. *Biometrika*, 8(1/2):250–254.

Quenouille, M. H. (1949). Approximate tests of correlation in time-series 3. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 45, pages 483–484. Cambridge University Press.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika*, 43(3/4):353–360.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, pages 832–837.

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, 90(432):1257–1270.

Scott, D. W. (2015). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Silverman, B. W. (2018). *Density estimation for statistics and data analysis*. Routledge.

Smirnov, N. V. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bull. Math. Univ. Moscou*, 2(2):3–14.

Sprent, P. and Smeeton, N. C. (2000). *Applied nonparametric statistical methods*. Chapman and Hall/CRC.

Stone, C. J. (1977). Consistent nonparametric regression. *The annals of statistics*, pages 595–620.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The annals of Statistics*, pages 1348–1360.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pages 1040–1053.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *the Annals of Statistics*, pages 1701–1728.

Tukey, J. (1958). Bias and confidence in not quite large samples. *Ann. Math. Statist.*, 29:614.

Wahba, G. (1990). *Spline models for observational data*, volume 59. Siam.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83.

Woodroofe, M. (1970). On choosing a delta-sequence. *The Annals of Mathematical Statistics*, 41(5):1665–1671.

CHAPTER 2. CONVERGENCE RATES FOR UNIFORM CONFIDENCE INTERVALS BASED ON LOCAL POLYNOMIAL REGRESSION ESTIMATORS

A Paper accepted by the Journal of Nonparametric Statistics (De Brabanter et al., 2016)

Abstract

We investigate the convergence rates of uniform bias corrected confidence intervals for a smooth curve using local polynomial regression for both the interior and boundary region. We discuss the cases when the degree of the polynomial is odd and even. The uniform confidence intervals are based on the volume-of-tube formula modified for biased estimators. We empirically show that the proposed uniform confidence intervals attain, at least approximately, nominal coverage. Finally, we investigate the performance of the volume-of-tube based confidence intervals for independent non Gaussian errors.

2.1 Introduction

Local polynomial regression is a well-known nonparametric regression technique in the area of statistics. It has found its roots in the late 1970s, see e.g. Stone (1977, 1980) and Cleveland (1979). This estimator gained in popularity due to its supremacy over local constant fitting i.e., Nadaraya-Watson kernel regression (Nadaraya, 1964; Watson, 1964). Attractive statistical properties and its easy implementation contributed further to its popularity (see the monograph of Fan and Gijbels (1996) for applications). Theoretical results can be found in the papers of Fan (1992, 1993) and Ruppert and Wand (1994).

Originally, uniform or simultaneous confidence intervals for a function m were constructed by studying the distribution of $\sup_{a\leq x\leq b}|\widehat{m}(x) - m(x)|$. A pioneering article relating to the study of the asymptotic distribution of density function estimates is due to Bickel and Rosenblatt (1973) and provides the basis for the derivation of the above mentioned asymptotic distribution. Extension to the nonparametric regression case is given by Knafl et al. (1985) and Härdle (1989) (for M-estimators). For the Priestley-Chao estimator, Eubank and Speckman (1993) used large scale upperbounds to develop uniform confidence intervals. Hall (1991) found that the convergence rates of these suprema are quite slow and hence confidence intervals based on this do not tend to perform well when considering small sample sizes. In order to overcome this drawback, bootstrap is usually the method of choice, see e.g. Neumann and Polzehl (1998) and Claeskens et al. (2003). But even then, some bootstrap procedures that are asymptotically correct can have poor finite sample performance (Loader, 1993).

Two problems arise in many approaches: bias of the nonparametric estimator and the smoothing parameter choice. To address the first problem, Hall (1992) used undersmoothing together with a pivotal statistic. If one would smooth less than the optimal amount then the bias will decrease asymptotically relative to the variance. Unfortunately, there does not seem to be a simple, practical rule for choosing just the right amount of undersmoothing (Neumann et al., 1995; Chen and Qin, 2000, 2002). Further, Hall (1992) showed that nonpivotal methods are not necessarily the most appropriate and pointed out some of the theoretical advantages of pivoting i.e., reduction of the error of the bootstrap distribution function estimate. Eubank and Speckman (1993) used an estimate of the second order derivative to correct for the bias of the nonparametric estimator. Xia (1998) extended the framework of Eubank and Speckman (1993) to local linear regression. In the two previous approaches the smoothing parameter was chosen in a data-driven way.

A more attractive and easy to implement method to construct uniform confidence bands is based on the volume-of-tube formula (Sun, 1993). The basis for this set-up dates back

to Weyl's formula (Weyl, 1939) for the volume of a tube of a manifold to the volume of the manifold embedded in the unit sphere. In the one dimensional case, the same result can also be obtained by the upcrossing theory of Rice (1939). A major advantage over upcrossing theory is that Weyl's formula can be extended to the multivariate case, see Sun et al. (1994). Although originally, the volume-of-tube formula assumed independent normally distributed errors, Loader and Sun (1997) empirically found that the volume-of-tube formula still holds, for unbiased estimators, without any modification for an error vector with a spherically symmetric distribution that has a density. They reported that the formula can break down in some extreme cases for discrete and highly skewed errors. It is also important to note that the justification for the volume-of-tube formula does not require that the sample size $n \to \infty$. Nevertheless, the choice of the bandwidth and handling biased estimators remains a critical task. Krivobokova et al. (2010) have successfully used the volume-of-tube formula to construct uniform confidence intervals for penalized splines and also treated the smoothing parameter as random. They showed that for the mixed-model representation of penalized splines no explicit bias correction is needed, assuming sufficient knots are taken so that the approximation bias is negligible. Unfortunately, this does not hold for the local polynomial regression case and therefore an explicit bias correction is needed.

In contrast to the frequentist setting, Bayesian confidence intervals (or credible intervals) are constructed based on the posterior distribution of the underlying process given the data. From a theoretical point of view this is optimal, but are hard to obtain. In general, Bayesian or credible intervals do not exactly coincide with frequentist confidence intervals for two reasons: First, credible intervals incorporate problem-specific contextual information from the prior distribution whereas confidence intervals are based only on the data. Second, credible intervals and confidence intervals treat smoothing parameters in very different ways (Bernardo and Smith, 2009). We refer the reader to Krivobokova et al. (2010) (and references therein) where the authors discuss the advantages of a mixed-model formulation which combines both frequentist and Bayesian approaches.

In this chapter we derive the asymptotic order of the width of bias-corrected uniform intervals, based on the volume-of-tube formula, for local polynomial regression of order p. The asymptotic order does not change from p even to its consecutive p odd degree in the interior region. An interesting fact is that the asymptotic order remains the same in the boundary region for p odd. In case p is even, the asymptotic order of the width of the intervals is slower in the boundary region than in the interior. The bandwidth is chosen in a usual way (cross-validation) and we treat the bandwidth as a stochastic variable in our results.

This chapter is organized as follows: In Section 2.2, we briefly describe the local polynomial regression framework and discuss the volume-of-tube formula. In Section 2.3, we derive our main theoretical results. We conduct a simulation study (with Gaussian and independent non Gaussian errors) and discuss two real life examples in Section 2.4 and Section 2.5 respectively. Proofs are given in Section 2.6. Finally, Section 2.7 states the conclusions and discusses directions for future research.

2.2 Local Polynomial Regression and Volume-of-tube Formula

2.2.1 Local polynomial regression

Consider the model

$$Y_i = m(X_i) + e_i, \qquad i = 1, \ldots, n, \tag{2.1}$$

where we observe an independent and identically distributed sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ distributed like (X, Y) with $\mathbf{E}[e|X] = 0$ and $\mathbf{E}[e^2|X] = \sigma_e^2 < \infty$. We want to estimate the regression curve $m(x) = \mathbf{E}[Y|X = x]$. Further, we assume that the errors are distributed as $N(0, \sigma_e^2)$. If m is a $p+1$ times continuously differentiable function, the local polynomial regression estimate of order p can be found by solving a weighted least squares problem yielding

$$\widehat{\beta} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y}$$

as the solution vector where $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\mathbf{W}$ is the $n \times n$ diagonal matrix of weights $\mathbf{W} = \text{diag}\{K_h(X_i - x_0)\}$ (with kernel K and bandwidth h) and

$$\mathbf{X} = \begin{pmatrix} 1 & (X_1 - x_0) & \cdots & (X_1 - x_0)^p \\ \vdots & \vdots & & \vdots \\ 1 & (X_n - x_0) & \cdots & (X_n - x_0)^p \end{pmatrix}.$$

The estimator $\widehat{\beta}_\nu$ in an arbitrary point x_0 is given by

$$\widehat{\beta}_\nu = \varepsilon_{\nu+1}^T \widehat{\beta} = \mathbf{l}_\nu(x_0)^T \mathbf{Y},$$

where $\varepsilon_{\nu+1}$=$(0, \ldots, 0, 1, 0, \ldots, 0)^T$ is a unit vector with a 1 on the $(\nu+1)$th position, $l_\nu(t) = \varepsilon_{\nu+1}^T (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \{1, th, \ldots, (th)^p\}^T K(t)/h$ and $\mathbf{l}_\nu(x_0) = (l_\nu\left(\frac{X_1 - x_0}{h}\right), \ldots, l_\nu\left(\frac{X_n - x_0}{h}\right))^T$. Note that $\widehat{\beta}_0 = \widehat{m}(x_0) = \mathbf{l}_0(x_0)^T \mathbf{Y}$ is the pth order local polynomial regression estimator.

2.2.2 Volume-of-tube formula and confidence intervals

The volume-of-tube formula dates back to Hotelling (1939) and Weyl (1939). Its use in uniform confidence intervals was developed by Naiman (1986, 1990), Sun et al. (1994) and others. Further theoretical results can be found in Sun (1993). In what follows we denote $\mathbb{X} = (X_1, \ldots, X_n)$.

Consider the regression model (2.1) and assume some unbiased linear smoother $\widehat{m}(x) = \mathbf{l}(x)^T \mathbf{Y}$ with $\mathbf{Var}[\widehat{m}(x)|\mathbb{X}] = \sigma_e^2 \| \mathbf{l}(x)\|^2$. Sun et al. (1994) showed that

$$\begin{aligned} \alpha &= \mathrm{P}\left[\sup_{x \in [a,b]} \frac{|\, \mathbf{l}(x)^T \mathbf{e}\,|}{\sigma_e \| \mathbf{l}(x)\|} \geq c\right] \\ &= \frac{\kappa_0}{\pi} \exp(-c^2/2) + 2(1 - \Phi(c)) + o\big(\exp(-c^2/2)\big), \end{aligned} \tag{2.2}$$

where $\kappa_0 = \int_a^b \left\| \frac{d}{dx} \frac{\mathbf{l}(x)}{\|\mathbf{l}(x)\|} \right\| dx$, α is the significance level, $\mathbf{e} = (e_1, \ldots, e_n)^T$ and $\Phi(\cdot)$ is the standard normal distribution. In case σ_e is unknown, one can use an estimate $\hat{\sigma}_e^2$ such that $\eta \widehat{\sigma}_e^2 / \sigma_e^2 \sim \chi_\eta^2$, then

$$\alpha \approx \frac{\kappa_0}{\pi} \left(1 + \frac{c^2}{\eta}\right)^{-\eta/2} + \mathrm{P}[|t_\eta| > c], \tag{2.3}$$

where χ^2_η is a χ^2-distributed random variable and t_η a t-distributed random variable both with η degrees of freedom. Any root solving algorithm (e.g. Newton-Raphson, Brent-Dekkers, etc.) can be used to determine c from (2.3). Then, a $100(1-\alpha)\%$ uniform confidence interval for $m(x)$ is given by

$$[\widehat{m}(x) - c\,\widehat{\sigma}_e \| \mathbf{l}(x)\|,\ \widehat{m}(x) + c\,\widehat{\sigma}_e \| \mathbf{l}(x)\|]. \tag{2.4}$$

Unfortunately, the latter confidence intervals will not work well for local polynomial regression since this is a biased estimator and consequently (2.2), (2.3) and (2.4) cannot be used. A seemingly attractive way of dealing with this problem is to shift the confidence intervals such that they account for the bias. These intervals are given by

$$[\widehat{m}(x) - (c + \widehat{\tau}(x))\,\widehat{\sigma}_e \| \mathbf{l}_0(x)\|,\ \widehat{m}(x) + (c - \widehat{\tau}(x))\,\widehat{\sigma}_e \| \mathbf{l}_0(x)\|]\,, \tag{2.5}$$

with

$$\widehat{\tau}(x) = \frac{\text{bias}[\widehat{m}(x)|\mathbb{X}]}{\sqrt{\mathbf{Var}[\widehat{m}(x)|\mathbb{X}]}}. \tag{2.6}$$

Just like the intervals (2.4), these intervals do not work well either. Assume that $\widehat{m}$ is a consistent estimate of m, subtracting the term $\widehat{\tau}(x)$ increases in general more variance than it reduces bias. Härdle and Marron (1991) and Sun et al. (1994) confirmed that such a correction fails. They observed that such a correction produces even worse results than no correction at all (see also Section 2.4).

A better way to proceed, is to consider the maximum standardized bias (Knafl et al., 1985; Hall and Titterington, 1988; Sun et al., 1994). Suppose the maximum of the function τ is known and let $\tau \in \mathcal{T}_\delta$ where

$$\mathcal{T}_\delta = \{\tau : \sup_{x\in[a,b]} |\tau(x)| \leq \delta\},$$

then Sun et al. (1994) showed that for classes of functions with bounded derivatives

$$\begin{aligned} \alpha &= \mathrm{P}\left[\sup_{x\in[a,b]} \frac{|\,\mathbf{l}_0(x)^T\,\mathbf{e} + \delta\,\text{bias}[\widehat{m}(x)|\mathbb{X}]|}{\sigma_e \| \mathbf{l}_0(x)\|} \geq c^\star\right] \\ &\leq \frac{\kappa_0}{2\pi}\left\{\mathrm{P}[\chi^2_2 > (c^\star - \delta)^2] + \mathrm{P}[\chi^2_2 > (c^\star + \delta)^2]\right\} + O\big(\delta^2\big) \end{aligned} \tag{2.7}$$

as $\delta \to 0$ where $c^\star$ is the critical value accounting for bias. Note that the final bias corrected formula in (4.8) Sun et al. (1994) has additional terms compared to (2.7). We chose to neglect these here since their simulation study suggested little is lost if these terms are ignored.

Although $\delta \to 0$ is needed for the volume-of-tube formula, we will provide empirical evidence to show that this condition can be relaxed. Our simulations (see Section 2.4) indicate that even by using cross-validation in determining an estimator for δ (see next paragraph) correct coverage can be obtained. It is, however an interesting theoretical challenge to relax this condition in the volume-of-tube formula.

2.2.3 Estimation of unknown quantities and numerical implementation

Before one can use (2.7), the maximum standardized bias δ needs to be estimated. According to (2.6), one needs an estimator for the bias of the local polynomial regression estimator. One possibility is to use the asymptotic expression for the bias of the local polynomial regression estimator. In this case, the leading term (higher order derivatives) has to be estimated. Such an approach was used in Eubank and Speckman (1993) and Xia (1998). Bias estimation, not relying on asymptotic expressions, is discussed in Müller (1985), Härdle et al. (1992) and Ruppert (1997). In what follows, we opt for the latter choice and hence the approximation will stay closer to the exact expression for the bias. A bias estimation method for local polynomial regression is described in [p. 113-114] Fan and Gijbels (1996) and will be our method of choice. For the sake of completeness, we briefly describe the idea. Using a Taylor expansion of order a, the bias of the local polynomial regression estimator in a point x_0 can be approximated by $(\mathbf{X}^T\,\mathbf{W}\,\mathbf{X})^{-1}\,\mathbf{X}^T\,\mathbf{W}\,\theta$, where θ is a $n \times 1$ vector with ith element

$$\beta_{p+1}(X_i - x_0)^{p+1} + \cdots + \beta_{p+a}(X_i - x_0)^{p+a}.$$

The optimal theoretical choice $a = 3$ entails that the bandwidth selector is $\sqrt{n}$-consistent in relative rate (Huang, 1995). However, to reduce the computational cost, we choose $a = 2$.

This will result in a bandwidth selector which is not far from being $\sqrt{n}$-consistent. The latter choice for a is also used in Fan and Gijbels (1996). Then the approximated bias can be estimated by

$$(\mathbf{X}^T\,\mathbf{W}\,\mathbf{X})^{-1}\begin{pmatrix}\widehat{\beta}_{p+1}S_{n,p+1}+\cdots+\widehat{\beta}_{p+a}S_{n,p+a}\\ \vdots \\ \widehat{\beta}_{2p+1}S_{n,p+1}+\cdots+\widehat{\beta}_{p+a}S_{n,2p+a}\end{pmatrix}, \tag{2.8}$$

where $\widehat{\beta}_{p+1},\ldots,\widehat{\beta}_{p+a}$ are the estimated coefficients obtained by fitting a polynomial degree $p+a$ locally. For this $(p+a)$th order fit on needs a bandwidth $h^\star$. For simplicity, cross-validation is used to determine the bandwidth $h^\star$ for this $(p+a)$th order fit, resulting into a bandwidth $\widehat{h^\star}$. Finally, the conditional bias of the regression estimate in a point x_0 is given by the first element of the estimated bias vector in (2.8).

The conditional variance of the local polynomial regression estimator in a point x, based on a bandwidth $\widehat{h}$ obtained by cross-validation, can be estimated as follows

$$\mathbf{Var}[\widehat{m}(x,\widehat{h})|\mathbb{X}] = \widehat{\sigma}_e^2\|\widehat{\mathbf{l}}_0(x)\|^2.$$

The error variance σ_e^2 can be estimated by means of the following model-free $\sqrt{n}$-consistent estimator (Hall et al., 1990)

$$\widehat{\sigma}_e^2 = \frac{1}{n-2}\sum_{i=1}^{n-2}(0.809Y_{[i]}-0.5Y_{[i+1]}-0.309Y_{[i+2]})^2, \tag{2.9}$$

where $Y_{[j]}$ denotes the jth order observation corresponding to the ordered $X_{[j]}$. Other model-free error variance estimators, not necessarily restricted to the one dimensional case, can be found in Devroye et al. (2013) and De Brabanter et al. (2014). An estimator for δ is then given by

$$\widehat{\delta} = \sup_{x\in[a,b]}|\widehat{\tau}(x)| = \sup_{x\in[a,b]}\left|\frac{\text{bias}[\widehat{m}(x,\widehat{h}^\star)|\mathbb{X}]}{\sqrt{\widehat{\sigma}_e^2\|\widehat{\mathbf{l}}_0(x)\|^2}}\right|. \tag{2.10}$$

A good approximation to $c^\star$ is obtained by solving, see (2.7), the following nonlinear equation

$$\begin{aligned} \alpha &\approx \frac{\kappa_0}{2\pi}\left\{\mathrm{P}[\chi_2^2 > (c^\star - \widehat{\delta})^2] + \mathrm{P}[\chi_2^2 > (c^\star + \widehat{\delta})^2]\right\} \\ &= \frac{\kappa_0}{2\pi}\left\{\exp\left(-\frac{(c^\star - \widehat{\delta})^2}{2}\right) + \exp\left(-\frac{(c^\star + \widehat{\delta})^2}{2}\right)\right\}. \end{aligned} \tag{2.11}$$

2.3 Main Theoretical Results

In practice, the bandwidth h is found in a data-driven way e.g., by cross-validation. Consequently, an estimate $\widehat{h}$ for the bandwidth is a random variable. Note that in the above procedure, the bandwidth was considered nonrandom. The following theorem shows that the variability due to bandwidth selection via cross-validation can be ignored in the local polynomial regression framework for n sufficiently large. Let $\widehat{\delta}$ and $\widehat{\sigma}_e^2$ be estimators for the maximum standardized bias δ and the error variance σ_e^2 respectively. The estimation procedures for both are given in (2.8) and (2.9). Before stating our main results, we state our assumptions:

(A1) m has bounded derivatives up to order $p+3$.

(A2) For the design points $X_i \in [a,b]$, $i = 1, \ldots, n$, there exist a density function f satisfying $0 < w \le f(x) \le W < \infty$ for some constants w and W. Further assume that f has a bounded derivative.

(A3) The kernel K is a symmetric density function and has bounded derivative K'. Further, the Fourier transformation of K is absolutely integrable.

(A4) $\mu_j = \int u^j K(u)\,du < \infty$ and $\nu_j = \int u^j K^2(u)\,du < \infty$ for $j = 0, \ldots, 2p$

(A5) $h \to 0$ and $nh \to \infty$ as $n \to \infty$, and $h = O\big(n^{-1/(2p+3)}\big)$

(A6) $h \to 0$ and $nh^3 \to \infty$ as $n \to \infty$, and $h = O\big(n^{-1/(2p+5)}\big)$

Theorem 2.3.1 *Under assumptions (A1)-(A5), let* $\widehat{\mathbf{l}}_0(x) = \mathbf{l}_0(x;\widehat{h})$, $\widehat{m}(x)$ *and* $\widehat{m}(x,\widehat{h})$ *be the estimator of the regression function based on the mean integrated squared error and cross-validation bandwidth respectively. Then, for a* $\sqrt{n}$*-consistent estimator of the error variance* σ_e^2*,*

$$\frac{\widehat{\mathbf{l}}_0(x)^T\,\mathbf{e} + \widehat{\delta}\,\mathrm{bias}[\widehat{m}(x,\widehat{h})|\mathbb{X}]}{\widehat{\sigma}_e\|\widehat{\mathbf{l}}_0(x)\|} = \frac{\mathbf{l}_0(x)^T\,\mathbf{e} + \delta\,\mathrm{bias}[\widehat{m}(x)|\mathbb{X}]}{\sigma_e\|\,\mathbf{l}_0(x)\|} + O_p\big(n^{-1/(4p+6)}\big)$$

for p *odd. Under assumptions (A1)-(A4) and assumption (A6)*

$$\frac{\widehat{\mathbf{l}}_0(x)^T\,\mathbf{e} + \widehat{\delta}\,\mathrm{bias}[\widehat{m}(x,\widehat{h})|\mathbb{X}]}{\widehat{\sigma}_e\|\widehat{\mathbf{l}}_0(x)\|} = \frac{\mathbf{l}_0(x)^T\,\mathbf{e} + \delta\,\mathrm{bias}[\widehat{m}(x)|\mathbb{X}]}{\sigma_e\|\,\mathbf{l}_0(x)\|} + O_p\big(n^{-1/(4p+10)}\big)$$

for p *even.*

The proof is given in Section 2.6. From Theorem 2.3.1 it immediately follows that smaller orders of p imply a smaller variability due to bandwidth selection.

Finally, bias-corrected confidence intervals for the local polynomial regression framework (i.e, $\nu = 0$) are given by

$$\left[\widehat{m}(x,\widehat{h}) - c^\star\,\widehat{\sigma}_e\|\widehat{\mathbf{l}}_0(x)\|,\ \widehat{m}(x,\widehat{h}) + c^\star\,\widehat{\sigma}_e\|\widehat{\mathbf{l}}_0(x)\|\right], \tag{2.12}$$

which will have approximately the desired coverage probability over classes of functions with bounded derivatives (Sun et al., 1994). Theorem 2.3.1 justifies replacing the bandwidth by its estimate obtained via cross-validation.

2.3.1 Asymptotic width of the confidence interval

The following theorem states the asymptotic width of the bias corrected confidence interval for local polynomial regression based on the volume-of-tube formula (2.7).

Theorem 2.3.2 *Under assumptions (A1)-(A5), the width of the bias corrected confidence interval* (2.12) *based on the volume-of-tube formula for local polynomial regression with* p *odd has asymptotic order* $O_p\big(\sqrt{\log n^{2/(2p+3)}}\,n^{-(p+1)/(2p+3)}\big)$*. Under assumptions (A1)-(A4) and assumption (A6), the width of the confidence interval has asymptotic order* $O_p\big(\sqrt{\log n^{2/(2p+5)}}\,n^{-(p+2)/(2p+5)}\big)$ *for* p *even.*

The proof is given in Section 2.6. Due to Theorem 2.3.1, Theorem 2.3.2 remains valid if the MISE bandwidth h is replaced by its estimate $\widehat{h}$ obtained by cross-validation. It immediately follows from Theorem 2.3.2 that the interval becomes narrower with increasing sample size n.

Up to a constant $\sqrt{2/(2p+3)}$ or $\sqrt{2/(2p+5)}$ the asymptotic width of the confidence interval (2.12) is similar to that of Eubank and Speckman (1993) for a twice continuously differentiable function. The asymptotic width also coincides with that of Krivobokova et al. (2010) for penalized splines. Note that the asymptotic width of the proposed confidence interval converges to zero only slightly slower than the optimal $((\log n)/n)^{2/5}$ rate of Hall and Titterington (1988) for a twice continuously differentiable function. Eubank and Speckman (1993) argued that the latter rates are obtained by choosing the bandwidth so that the width of the confidence interval is minimized. Eubank and Speckman (1993) stressed that the optimal rate of decay from Hall and Titterington (1988) for the width of the confidence interval cannot be obtained by using a bandwidth of order $O\big(n^{-1/(2p+3)}\big)$ or $O\big(n^{-1/(2p+5)}\big)$ for p odd and p even respectively. Further, they conjectured that their rate of $\sqrt{\log n}\, n^{-2/5}$ is the best attainable when the bandwidth is chosen to minimize the mean squared error. In the local polynomial framework, narrower intervals can be obtained not only by increasing the sample size n but also by choosing a larger bandwidth h.

In the proof of Theorem 2.3.2 (see Section 2.6), we obtain an asymptotic expression for the length of the manifold $\mathcal{M} = \{\mathbf{l}_0(x)/\|\,\mathbf{l}_0(x)\|; x \in \mathbb{R}\}$ given by

$$\kappa_0 = \int_a^b \left\| \frac{d}{dx} \frac{\mathbf{l}_0(x)}{\|\,\mathbf{l}_0(x)\|} \right\| dx = \frac{b-a}{h} \sqrt{\left[\frac{\int K_{1,p}^{\star 2}(t)\,dt}{\int K_{0,p}^{\star 2}(t)\,dt} - \frac{\left[\int K_{0,p}^{\star}(t) K_{1,p}^{\star T}(t)\,dt\right]^2}{\left[\int K_{0,p}^{\star 2}(t)\,dt\right]^2} \right]} \{1 + o_p(1)\},$$

where $K_{\nu,p}^{\star}(t) = \varepsilon_{\nu+1}^T S^{-1}(1, t, \ldots, t^p)^T K(t)$, with $S = (\mu_{j+l})_{0 \le j,l \le p}$, denotes the equivalent kernel in a point t. By the Cauchy-Schwartz inequality, it immediately follows that

$$\frac{\left[\int K_{0,p}^{\star}(t) K_{1,p}^{\star T}(t)\,dt\right]^2}{\left[\int K_{0,p}^{\star 2}(t)\,dt\right]^2} \le \frac{\int K_{1,p}^{\star 2}(t)\,dt}{\int K_{0,p}^{\star 2}(t)\,dt}.$$

In fact, little is lost if we ignore the second term under the square root. It is clear that the length of the manifold $\mathcal{M}$ is determined by n (and hence h), range of the data, p and the kernel function. For same n, increasing p will decrease the length of the manifold.

Another interesting fact is that the asymptotic order does not change from p even to its consecutive p odd degree. For example, for local constant regression ($p = 0$) and local linear regression ($p = 1$) the asymptotic order of the confidence intervals are both equal to $O_p\big(\sqrt{\log n^{2/5}}\, n^{-2/5}\big)$. As we will illustrate next, this rate can only be achieved in the interior region and not in the boundary region.

2.3.2 Asymptotic rate at the boundaries

The asymptotic width of the proposed confidence interval (see Theorem 2.3.2) has been derived assuming that the points under consideration belong to the interior region. However, when points belong to the boundary region, there are differences in the bias and variance of the local polynomial regression smoother. Without loss of generality, assume that the density has bounded support on $[0, 1]$. A left boundary point is of the form $x = ch$ with $c \geq 0$, whereas a right boundary point is of the form $x = 1 - ch$. Then for the left boundary, besides the fact that all the integration limits in the proofs (see Section 2.6) will change from $[-\infty, +\infty]$ to $[-c, +\infty]$ and f will be evaluated in $0+$ (i.e., zero approaching from the right), the asymptotic order of the variance will remain the same. The major difference is in the asymptotic order of the bias for p even (Fan and Gijbels, 1992; Ruppert and Wand, 1994). In the interior the bias is of order $O_p\big(h^{p+2}\big)$, while in the boundary the order is $O_p\big(h^{p+1}\big)$. Consequently, in the boundary region, this will change the result of Theorem 2.3.2 for p even. For p odd and in the boundary, Theorem 2.3.2 remains valid. Theorem 2.3.3 states the asymptotic width of the bias-corrected confidence interval at the boundaries for p even. In the latter, it turns out that the rate of the confidence interval is also determined by the order of the maximum standardized bias (see proof of Theorem 2.3.3).

Theorem 2.3.3 *Under assumptions (A1)-(A3), (A6) and* $\mu_{j,c} = \int_{-c}^{+\infty} u^j K(u)\,du < \infty$ *and* $\nu_{j,c} = \int_{-c}^{+\infty} u^j K^2(u)\,du < \infty$ *for* $j = 0, \ldots, 2p$*, the width of the bias corrected confidence interval* (2.12) *based on the volume-of-tube formula for local polynomial regression at the left boundary has asymptotic order* $O_p\big(\sqrt{n^{2/(2p+5)}}\, n^{-(p+2)/(2p+5)}\big)$ *for* p *even.*

The proof is given in Section 2.6. Under assumptions (A1)-(A3), (A6) and $\int_{-\infty}^{c} u^j K(u)\,du < \infty$ and $\int_{-\infty}^{c} u^j K^2(u)\,du < \infty$ for $j = 0, \ldots, 2p$, the asymptotic order of the confidence interval at the right boundary is exactly the same. Hence for local constant regression ($p = 0$) and local linear regression ($p = 1$) the asymptotic order of the width of the confidence intervals in the boundary region are $O_p\big(\sqrt{n^{2/5}}\, n^{-2/5}\big)$ and $O_p\big(\sqrt{\log n^{2/5}}\, n^{-2/5}\big)$ respectively. The asymptotic order of the confidence interval based on the local linear regression estimator achieves a faster rate than local constant fitting in the boundary region. Therefore, odd degrees are to be preferred.

2.4 Simulations

2.4.1 Normal errors

To verify our theoretical results, we conduct the following simulation study. Consider the two functions used in Xia (1998), Eubank and Speckman (1993) and Wahba (1983):

$$m_1(X) = \sin^2\{2\pi(X - 0.5)\} \qquad \text{(Model I)}$$

and

$$m_2(X) = \frac{6}{10}\beta_{30,17}(X) + \frac{4}{10}\beta_{3,11}(X) \qquad \text{(Model II)},$$

with $\beta_{l,m}(X) = \Gamma(l+m)\{\Gamma(l)\Gamma(m)\}^{-1} X^{l-1}(1-X)^{m-1}$. The X values were taken to be uniformly distributed on $[0, 1]$. Sample sizes $n = 100$, 250 and 500 were used and the errors were taken to be independent $N(0, \sigma_e^2)$ distributed with $\sigma_e = 0.05, 0.1, 0.3$ and $\sigma_e = 0.5$. In our analysis, we have used the Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}} \exp(-u^2/2)$ for the local polynomial regression estimator of order 1, 2 and 3. The bandwidths (for regression

estimation and bias estimation) were chosen via the cross-validation routine implemented in the R package *locpol* (Ojeda Cabrera, 2012). We generated 10,000 replicate samples for each experimental setting, using a different seed for every case. The results for the 95% uniform confidence intervals are reported in Table 2.1. Four types of confidence intervals are considered i.e., bias-corrected confidence intervals (2.12) with estimated values for δ, see (2.10), range between 0.35 and 0.98 for Model I and between 0.2 and 1.28 for Model II (for all values of p considered), confidence intervals proposed by Xia (1998), shifting the confidence interval to account for bias (2.5) and ignoring the bias (2.4).

The rows labeled as (2.12) represent the coverage probabilities and corresponding areas for bias-corrected confidence intervals (2.12). They should result in a coverage probability close to the nominal level, which is indeed the case. Only for low noise levels, these intervals are conservative. As expected from our results in Theorem 2.3.2, for a fixed p, the width of the intervals (and therefore the areas) become smaller for increasing sample sizes. This simulation also shows, for a fixed sample size, there is a decrease of the width of the confidence interval going from an even to an odd p. In the transition from an odd to an even p this difference is negligible (see Theorem 2.3.2).

A second type of confidence intervals are the ones proposed by Xia (1998) for local linear regression. These confidence intervals are based on large sample upperbounds and bias correction was performed by estimating the leading order term in the bias of the local linear estimator i.e., estimation of the second order derivative of the regression function. In order to estimate the second order derivative we used local cubic fitting. We first used cross-validation to obtain a bandwidth for the local cubic fit. Then, we made an adjustment to the cross-validation bandwidth to obtain a suitable bandwidth $\widehat{b}$ for the estimation of the second order derivative. The latter has also been done in Fan and Gijbels (1995) and [p. 67]Fan and Gijbels (1996). These confidence intervals are given by

$$\widehat{m}(x,\widehat{h}) - \frac{1}{2}\widehat{m}''(x,\widehat{b})\widehat{h}^2 \pm L_\alpha(x), \tag{2.13}$$

where

$$L_\alpha(x) = \widehat{\sigma}_e \widehat{V} \frac{\left[\sqrt{-2\log(\widehat{h})} + \frac{A-\chi_\alpha}{\sqrt{-2\log(\widehat{h})}}\right]}{\left[\sum_{i=1}^n K\left(\frac{X_i-x}{\widehat{h}}\right)\left(s_{n,\widehat{h},2} - \frac{X_i-x}{\widehat{h}} s_{n,\widehat{h},1}\right)\right]^{1/4}}, \quad A = \log\left[\frac{1}{2\pi}\sqrt{\frac{\int K'(u)^2\,du}{\int K^2(u)\,du}}\right],$$

with $\chi_\alpha = \log[-\log(1-\alpha)/2]$ and $s_{n,\widehat{h},j} = \sum_{i=1}^n K\left(\frac{X_i-x}{\widehat{h}}\right)\left(\frac{X_i-x}{\widehat{h}}\right)^j$. Finally, to improve accuracy for finite samples we choose

$$\widehat{V}^2 = \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^n \frac{1}{n\widehat{h}\widehat{f}^3(X_j)}[\widehat{f}(X_j) - \widehat{f}'(X_j)(X_i - X_j)]^2 K^2\left(\frac{X_i - X_j}{\widehat{h}}\right).$$

The density $\widehat{f}$ and its derivative $\widehat{f}'$ were estimated with the Parzen-Rosenblatt kernel density estimator. Both bandwidths were chosen according to the normal reference rule (Silverman, 2018, p. 48). These confidence bands seem to perform quite well when the noise level is low, confirming the results of Xia (1998). However, for larger noise levels they lead to severe undercoverage. In general these intervals tend to be wider than the ones based on the volume-of-tube formula.

Both intervals (2.4) and (2.5) lead to severe undercoverage when $\sigma_e \geq 0.1$. Overall, these intervals have on average 5%-10% less coverage compared to nominal coverage for all combinations of n and p when $\sigma_e \geq 0.1$. These results coincide with the findings of Härdle and Marron (1991) and Sun et al. (1994).

2.4.2 Independent non Gaussian errors

Unfortunately, it is not true that results (2.3) and (2.7) for the Gaussian error case carries through trivially to any symmetric error distribution even if the error variance σ_e^2 is known. If the error variance σ_e^2 is unknown, Loader and Sun (1997) empirically showed that the volume-of-tube formula is robust for the spherically symmetric error distributions in case of unbiased estimators. In what follows, we will empirically investigate the performance of the volume-of-tube formula in case of a biased estimator. Consider the following function

$$m_3(X) = 32\exp\{-8(1-2X)^2\}(1-2X) \tag{2.14}$$

Table 2.1: Coverage probabilities and median areas for models with normal errors

Type	σ_e	p	**Model I** Coverages and (area) for different n			**Model II** Coverages and (area) for different n		
			$n = 100$	$n = 250$	$n = 500$	$n = 100$	$n = 250$	$n = 500$
(2.12)	0.05	1	0.99 (0.21)	0.98 (0.12)	0.97 (0.08)	1 (0.58)	0.99 (0.20)	0.97 (0.08)
	0.05	2	0.99 (0.15)	0.96 (0.08)	0.95 (0.06)	1 (0.42)	0.99 (0.14)	0.95 (0.06)
	0.05	3	0.99 (0.14)	0.95 (0.07)	0.94 (0.06)	1 (0.42)	0.99 (0.14)	0.94 (0.05)
(2.13)	0.05	1	0.99 (0.20)	0.96 (0.10)	0.95 (0.08)	1 (0.61)	0.99 (0.21)	0.99 (0.12)
(2.5)	0.05	1	1 (0.19)	0.96 (0.10)	0.94 (0.07)	1 (0.61)	1 (0.20)	1 (0.61)
	0.05	2	1 (0.15)	0.96 (0.07)	0.93 (0.05)	1 (0.44)	1 (0.14)	1 (0.43)
	0.05	3	1 (0.15)	0.96 (0.070	0.93 (0.05)	1 (0.44)	1 (0.14)	1 (0.44)
(2.4)	0.05	1	1 (0.19)	0.95 (0.10)	0.90 (0.16)	1 (0.61)	1 (0.20)	0.97 (0.11)
	0.05	2	1 (0.15)	0.85 (0.08)	0.89 (0.14)	1 (0.43)	1 (0.14)	0.96 (0.08)
	0.05	3	1 (0.15)	0.84 (0.08)	0.88 (0.14)	1 (0.44)	1 (0.14)	0.97 (0.08)
(2.12)	0.1	1	0.98 (0.29)	0.97 (0.19)	0.97 (0.14)	1 (0.55)	0.96 (0.25)	0.94 (0.17)
	0.1	2	0.96 (0.22)	0.94 (0.14)	0.94 (0.10)	0.99 (0.43)	0.96 (0.19)	0.94 (0.14)
	0.1	3	0.95 (0.21)	0.94 (0.13)	0.95 (0.09)	0.99 (0.43)	0.97 (0.20)	0.94 (0.13)
(2.13)	0.1	1	0.95 (0.26)	0.91 (0.17)	0.92 (0.13)	0.99 (0.58)	0.98 (0.27)	0.95 (0.18)
(2.5)	0.1	1	0.97 (0.26)	0.92 (0.16)	0.91 (0.12)	1 (0.57)	0.97 (0.25)	0.90 (0.17)
	0.1	2	0.92 (0.21)	0.82 (0.13)	0.80 (0.09)	1 (0.45)	0.96 (0.19)	0.88 (0.13)
	0.1	3	0.96 (0.22)	0.92 (0.13)	0.92 (0.09)	0.99 (0.45)	0.97 (0.19)	0.93 (0.14)
(2.4)	0.1	1	0.96 (0.25)	0.90 (0.16)	0.88 (0.13)	1 (0.57)	0.97 (0.25)	0.89 (0.17)
	0.1	2	0.95 (0.21)	0.89 (0.13)	0.87 (0.10)	1 (0.45)	0.96 (0.19)	0.87 (0.13)
	0.1	3	0.97 (0.22)	0.92 (0.13)	0.92 (0.10)	1 (0.45)	0.97 (0.20)	0.93 (0.14)
(2.12)	0.3	1	0.95 (0.60)	0.96 (0.42)	0.96 (0.32)	0.98 (0.96)	0.98 (0.65)	0.98 (0.49)
	0.3	2	0.95 (0.58)	0.96 (0.40)	0.97 (0.30)	0.96 (0.78)	0.95 (0.50)	0.96 (0.37)
	0.3	3	0.94 (0.57)	0.96 (0.39)	0.97 (0.28)	0.95 (0.80)	0.95 (0.51)	0.96 (0.38)
(2.13)	0.3	1	0.78 (0.54)	0.80 (0.36)	0.84 (0.26)	0.92 (0.80)	0.87 (0.51)	0.88 (0.40)
(2.5)	0.3	1	0.86 (0.53)	0.86 (0.37)	0.87 (0.28)	0.95 (0.81)	0.90 (0.52)	0.89 (0.39)
	0.3	2	0.82 (0.48)	0.82 (0.32)	0.84 (0.24)	0.91 (0.71)	0.83 (0.44)	0.82 (0.32)
	0.3	3	0.88 (0.51)	0.89 (0.34)	0.88 (0.25)	0.93 (0.75)	0.89 (0.47)	0.86 (0.34)
(2.4)	0.3	1	0.87 (0.53)	0.87 (0.37)	0.87 (0.28)	0.92 (0.81)	0.86 (0.52)	0.85 (0.39)
	0.3	2	0.86 (0.48)	0.86 (0.32)	0.86 (0.24)	0.92 (0.71)	0.85 (0.44)	0.83 (0.32)
	0.3	3	0.88 (0.51)	0.90 (0.34)	0.90 (0.25)	0.91 (0.75)	0.88 (0.47)	0.89 (0.34)
(2.12)	0.5	1	0.93 (1.03)	0.94 (0.70)	0.95 (0.54)	0.95 (1.31)	0.96 (0.93)	0.98 (0.72)
	0.5	2	0.94 (1.01)	0.96 (0.68)	0.96 (0.52)	0.93 (1.12)	0.93 (0.78)	0.94 (0.58)
	0.5	3	0.94 (1.02)	0.95 (0.67)	0.96 (0.52)	0.93 (1.11)	0.93 (0.78)	0.95 (0.58)
(2.13)	0.5	1	0.56 (0.81)	0.73 (0.55)	0.77 (0.42)	0.80 (1.14)	0.80 (0.78)	0.84 (0.59)
(2.5)	0.5	1	0.81 (0.75)	0.83 (0.54)	0.84 (0.41)	0.88 (1.12)	0.85 (0.75)	0.86 (0.57)
	0.5	2	0.77 (0.73)	0.80 (0.50)	0.81 (0.37)	0.81 (1.01)	0.79 (0.66)	0.80 (0.49)
	0.5	3	0.84 (0.77)	0.86 (0.52)	0.87 (0.39)	0.85 (1.07)	0.85 (0.71)	0.86 (0.53)
(2.4)	0.5	1	0.83 (0.77)	0.86 (0.54)	0.86 (0.41)	0.86 (1.12)	0.84 (0.75)	0.84 (0.57)
	0.5	2	0.82 (0.73)	0.84 (0.49)	0.85 (0.37)	0.83 (1.01)	0.83 (0.69)	0.83 (0.49)
	0.5	3	0.84 (0.78)	0.87 (0.52)	0.89 (0.39)	0.84 (1.07)	0.86 (0.71)	0.87 (0.52)

Coverage probabilities and median areas (between parenthesis) for Model I and II with nominal level 0.95 using bias-corrected confidence intervals (2.12), confidence bands (2.13), shifting the confidence intervals (2.5) and ignoring the bias (2.4) for $p = 1, 2, 3$ and four different σ_e. Results are based on 10,000 repetitions using a different seed for every case.

where the X values were taken to be uniformly distributed on $[0, 1]$. The errors were taken to be Laplace distributed with location parameter set to zero and varying scale parameter b. As before, estimated values for δ were obtained by (2.10). In a second example, the errors have an exponential distribution (with varying parameter λ), shifted to have zero mean. Further, we used the same settings as in the previous paragraph. Table 2.2 and Table 2.3 display the coverage probabilities and corresponding median area (between parenthesis) based on 1,000 replicate samples for each experimental setting for local linear and local cubic regression with Laplacian and exponential errors respectively for the confidence intervals (2.12). In the case of small sample sizes, the achieved coverage is not close to the nominal. For a small value of the parameter λ (exponential errors), the bias-corrected confidence intervals (2.12) lead to severe undercoverage. This is not surprising considering the severity of the errors. The conclusions from these examples are two-fold. First, heavy tails and skewness seem to result in coverage probabilities close to nominal for medium and large sample sizes. Second, there does not seem to be a large difference between the local linear and cubic fits in terms of coverage probabilities. The conclusions of Theorem 2.3.2 still seem to hold for independent non Gaussian errors.

Table 2.2: Coverage probabilities and median areas (between parenthesis) for the regression function (2.14) with nominal level 0.95 using (2.12) for Laplacian errors for varying scale parameter b.

b	p	Coverages and (area) for different n		
		$n = 100$	$n = 250$	$n = 500$
0.1	1	1 (1.08)	0.99 (0.47)	0.97 (0.31)
0.1	3	1 (0.81)	0.99 (0.32)	0.98 (0.21)
0.3	1	0.99 (1.44)	0.96 (0.90)	0.96 (0.68)
0.3	3	0.98 (1.16)	0.96 (0.70)	0.95 (0.50)
0.5	1	0.95 (1.93)	0.96 (1.31)	0.96 (0.99)
0.5	3	0.95 (1.64)	0.95 (1.05)	0.95 (0.78)
0.8	1	0.94 (2.63)	0.95 (1.85)	0.96 (1.42)
0.8	3	0.92 (2.36)	0.94 (1.57)	0.95 (1.16)

Table 2.3: Coverage probabilities and median areas (between parenthesis) for the regression function (2.14) with nominal level 0.95 using (2.12) for exponential errors for varying parameter λ.

λ	p	Coverages and (area) for different n		
		$n = 100$	$n = 250$	$n = 500$
0.1	1	0.61 (11.68)	0.76 (8.90)	0.90 (7.28)
0.1	3	0.90 (14.24)	0.84 (10.0)	0.87 (7.47)
0.3	1	0.86 (5.71)	0.93 (4.16)	0.96 (3.21)
0.3	3	0.86 (5.68)	0.94 (3.90)	0.95 (2.94)
0.5	1	0.92 (3.97)	0.94 (2.83)	0.95 (2.19)
0.5	3	0.92 (3.72)	0.94 (2.54)	0.95 (1.91)
0.8	1	0.93 (2.82)	0.95 (1.99)	0.95 (1.53)
0.8	3	0.94 (2.52)	0.94 (1.70)	0.95 (1.27)

In a final toy example, we investigate the performance of the volume-of-tube for the discrete point distribution case e.g. let the errors have the following distribution: $\mathrm{P}[e_i = -1] = \gamma$ and $\mathrm{P}[e_i = \gamma/(1-\gamma)] = 1-\gamma$, for values of γ ranging from 0.3 to 0.95 (see Table 2.4). Although the volume-of-tube formula requires the existence of a density, the proposed methodology works quite well for this two point discrete distribution. Only for values $\gamma \geq 0.9$, the achieved coverage is less than the nominal.

2.5 Examples

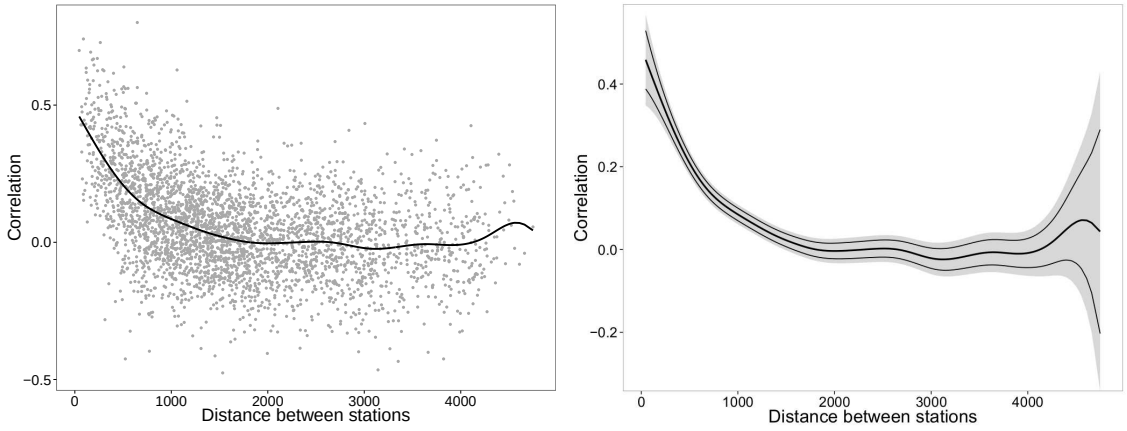
In this section we present two real life examples to illustrate the bias-corrected uniform confidence intervals. In a first example, we consider the sulfate wet depositions data set (Oehlert, 1993). The data set is freely available at http://people.stern.nyu.edu/jsimonof/SmoothMeth/Data/Tab/sulfate.tab. It describes the relationship of distance (measured in kilometers) between 3321 pairs of measuring stations and the correlation of adjusted sulfate wet deposition levels. All observations are shown in Figure 2.1a. The gray area in Figure 2.1b shows the 95% uniform confidence interval (2.12) for the local linear regression estimate i.e., $p = 1$. The volume-of-tube yielded the critical value $c^\star = 4.787$ for a bandwidth $\widehat{h} = 190.624$ obtained by cross-validation. The two thin lines represent the

Table 2.4: Coverage probabilities and median areas (between parenthesis) for the regression function (2.14) with nominal level 0.95 using (2.12) for the two point discrete distribution errors for varying parameter γ.

γ	p	Coverages and (area) for different n		
		$n = 100$	$n = 250$	$n = 500$
0.3	1	0.99 (1.85)	0.98 (1.23)	0.98 (0.94)
0.3	3	0.97 (1.54)	0.97 (0.98)	0.96 (0.72)
0.5	1	0.98 (2.43)	0.97 (1.69)	0.98 (1.29)
0.5	3	0.95 (2.13)	0.96 (1.40)	0.96 (1.04)
0.7	1	0.96 (3.27)	0.96 (2.30)	0.97 (1.78)
0.7	3	0.94 (2.99)	0.95 (2.02)	0.95 (1.49)
0.9	1	0.85 (5.36)	0.91 (3.86)	0.94 (2.97)
0.9	3	0.84 (5.35)	0.88 (3.59)	0.94 (2.73)
0.95	1	0.77 (6.85)	0.88 (5.05)	0.91 (3.93)
0.95	3	0.85 (6.25)	0.88 (4.99)	0.90 (3.68)

95% uniform confidence intervals ignoring the fact that the smoother is biased. The critical value $c = 3.059$ of the volume-of-tube is obtained by (2.3) with $\eta = n - \text{tr}(\mathbf{L}) = 3309.586$, where $\mathbf{L}$ denotes the smoother matrix of the local linear regression estimate. The former confidence bands could be used to perform a test for significance of bumps and dips e.g. in collaboration with SiZer (Chaudhuri and Marron, 1999), between 2000 and 4000 kilometers.

The second example involves the age and income data for a sample of 205 Canadian workers from a 1971 Canadian Census Public Use Tape, all of whom were educated to grade 13 (Ullah, 1985). The data set is part of the R package *SemiPar* (Ruppert et al., 2003). Figure 2.2 illustrates the bias-corrected uniform confidence interval (2.12) versus shifting the confidence interval to account for bias (2.5) for local linear and local cubic regression respectively. The bandwidths for both estimators were obtained by cross-validation and were $\widehat{h} = 3.268$ and $\widehat{h} = 8.773$ for the local linear and local cubic regression respectively. The critical values, for the intervals (2.12), obtained by the volume-of-tube formula (2.7) were $c^\star = 4.46$ and $c^\star = 4.069$ for the local linear and local cubic case respectively. The

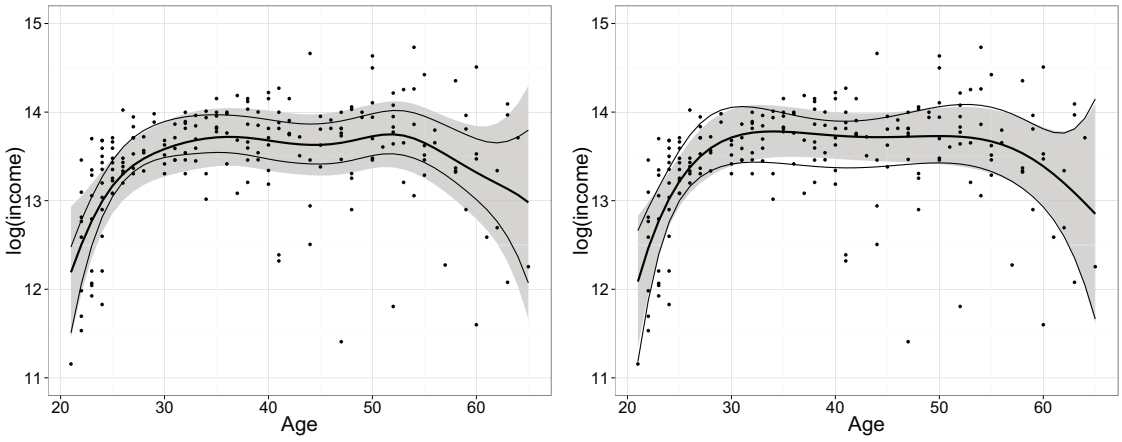


(a) (b)

Figure 2.1: Sulfate wet depositions data. (a) The data and a local linear regression fit with bandwidth $\widehat{h} = 190.624$ obtained by cross-validation. (b) 95% uniform confidence intervals when taking the bias into account (grey area) and when ignoring the bias (full lines).

area of the bias-corrected confidence intervals are 36.52 and 32.36 for local linear and local cubic respectively which coincides with the findings in Theorem 2.3.2. In case the intervals were shifted to account for bias, critical values were obtained from (2.3) yielding $c = 2.92$ and $c = 2.84$, with corresponding areas 23.92 and 22.63, for the local linear and local cubic case respectively.

It is interesting to see that for local linear regression, there is a dip around the age of 45. This dip is not present when a local cubic fit is used. From a socio-economic perspective this could be interesting and worth further investigation. Eventually, the regression function reaches a plateau between ages 33 and 53. The latter observation suggests that middle-age workers do not have much (if any) of a raise. As reported in Ruppert et al. (2003), these data are cross-sectional and not longitudinal. Therefore, one must be very cautious about interpreting cross-sectional data in such a longitudinal fashion. As in the previous example, SiZer can help to detect whether such a dip is really there or not.



(a) (b)

Figure 2.2: Age and log(income) data. (a) The data and a local linear regression fit (bold line) with bandwidth $\widehat{h} = 3.268$ obtained by cross-validation with corresponding 95% uniform confidence interval (gray area) and shifting the confidence interval (full lines). (b) Similar as (a) when fitting a local cubic fit with $\widehat{h} = 8.77$ obtained by cross-validation.

2.6 Proofs

Proof of Theorem 2.3.1. Consider the following decomposition

$$\begin{aligned}
&\frac{\widehat{\mathbf{l}}_0(x)^T\,\mathbf{e} + \widehat{\delta}\,\mathrm{bias}[\widehat{m}(x,\widehat{h})|\mathbb{X}]}{\widehat{\sigma}_e\|\widehat{\mathbf{l}}_0(x)\|} \\
&= \left[\frac{\mathbf{l}_0(x)^T\,\mathbf{e} + \delta\,\mathrm{bias}[\widehat{m}(x)|\mathbb{X}]}{\sigma_e\|\,\mathbf{l}_0(x)\|} + \frac{[\widehat{\mathbf{l}}_0(x)^T - \mathbf{l}_0(x)^T]\,\mathbf{e} - \delta\,\mathrm{bias}[\widehat{m}(x)|\mathbb{X}] + \widehat{\delta}\,\mathrm{bias}[\widehat{m}(x,\widehat{h})|\mathbb{X}]}{\sigma_e\|\,\mathbf{l}_0(x)\|}\right]\frac{\|\,\mathbf{l}_0(x)\|}{\|\widehat{\mathbf{l}}_0(x)\|}\frac{\sigma_e}{\widehat{\sigma}_e}.
\end{aligned}$$

In what follows, we derive the asymptotic order for the second term in the above equation in case p is odd. The proof for p even is completely analogous. Under assumptions (A2)-(A3), m has bounded derivatives up to order $p+2$ and for p odd, Xia and Li (2002) showed that the mean integrated squared error (MISE) optimal bandwidth h and the cross-validated bandwidth $\widehat{h}$ satisfy the following relation

$$\widehat{h} = h + O_p\big(n^{-3/(4p+6)}\big). \tag{2.15}$$

Using [Theorem 3.1] Fan and Gijbels (1996), assumption (A5) and (2.15), we have that

$$\begin{aligned}\widehat{\mathbf{l}}_0(x)^T\,\mathbf{m} - \mathbf{l}_0(x)^T\,\mathbf{m} &= \big(\mathbf{E}[\widehat{m}(x,\widehat{h})|\mathbb{X}] - m(x)\big) - \big(\mathbf{E}[\widehat{m}(x)|\mathbb{X}] - m(x)\big)\\ &= \text{bias}[\widehat{m}(x,\widehat{h})|\mathbb{X}] - \text{bias}[\widehat{m}(x)|\mathbb{X}]\\ &= O_p\big(\widehat{h}^{p+1} - h^{p+1}\big) = O_p\big(n^{-1/2}\big).\end{aligned}$$

with $\mathbf{m} = (m(X_1),\ldots,m(X_n))^T$. Next, denote the matrices $S = (\mu_{j+l})_{0\le j,l\le p}$ and $S^\star = (\nu_{j+l})_{0\le j,l\le p}$, then under assumption (A4)

$$\|\mathbf{l}_0(x)\|^2 = \frac{1}{nhf(x)}\varepsilon_1^T S^{-1}S^\star S^{-1}\varepsilon_1\{1+o_p(1)\}$$

with $\varepsilon_1 = (1,0,0,\ldots,0)^T$ a unit vector with 1 on the first position. Under assumption (A5), it follows that

$$\begin{aligned}\frac{\|\,\mathbf{l}_0(x)\|^2}{\|\widehat{\mathbf{l}}_0(x)\|^2} &= 1 + \frac{\|\,\mathbf{l}_0(x)\|^2 - \|\widehat{\mathbf{l}}_0(x)\|^2}{\|\widehat{\mathbf{l}}_0(x)\|^2}\\ &= 1 + O_p\big((\widehat{h}-h)/h\big)\\ &= 1 + O_p\big(n^{-1/(4p+6)}\big).\end{aligned}$$

Finally, using the fact that $\delta = O_p(1)$,

$$[\widehat{\mathbf{l}}_0(x)^T - \mathbf{l}_0(x)^T]\,\mathbf{e} = [\widehat{\mathbf{l}}_0(x)^T - \mathbf{l}_0(x)^T]\{\mathbf{m}+O_p(1)\} - [\widehat{\mathbf{l}}_0(x)^T - \mathbf{l}_0(x)^T]\,\mathbf{m},$$

and $\widehat{\sigma}_e$ is a $\sqrt{n}$-consistent estimator for σ_e, it immediately follows that

$$\begin{aligned}&\frac{\widehat{\mathbf{l}}_0(x)^T\,\mathbf{e} + \widehat{\delta}\,\text{bias}[\widehat{m}(x,\widehat{h})|\mathbb{X}]}{\widehat{\sigma}_e\|\widehat{\mathbf{l}}_0(x)\|}\\ =\;& \left[\frac{\mathbf{l}_0(x)^T\,\mathbf{e} + \delta\,\text{bias}[\widehat{m}(x)|\mathbb{X}]}{\sigma_e\|\,\mathbf{l}_0(x)\|} + O_p\big(n^{-1/(4p+6)}\big)\right]\big\{1+O_p\big(n^{-1/(4p+6)}\big)\big\}\big\{1-O_p\big(n^{-1/2}\big)\big\}\\ =\;& \frac{\mathbf{l}_0(x)^T\,\mathbf{e} + \delta\,\text{bias}[\widehat{m}(x)|\mathbb{X}]}{\sigma_e\|\,\mathbf{l}_0(x)\|} + O_p\big(n^{-1/(4p+6)}\big).\end{aligned}$$

Proof of Theorem 2.3.2. The width of the confidence intervals based on the volume-of-tube formula for local polynomial regression at a fixed x is determined by the critical value $c^\star$ and estimated standard deviation $\widehat{\sigma}_e\|\widehat{\mathbf{l}}_0(x)\|$. From (2.7), it follows that

$c^\star = \sqrt{\log[\kappa_0{}^2\{1+O_p(1)\}]} + \widehat{\delta}$, where $O_p(1)$ is bounded in probability for $c^\star \to \infty$. Based on a $\sqrt{n}$-consistent variance estimator, using the fact that $\| \mathbf{l}_0(x)\|^2 \|\widehat{\mathbf{l}}_0(x)\|^{-2} = 1 + O_p\big(n^{-1/(4p+6)}\big)$ (see proof of Theorem 2.3.1) and p odd, we have

$$\widehat{\sigma}_e \|\widehat{\mathbf{l}}_0(x)\| = O_p\big(n^{-(p+1)/(2p+3)}\big).$$

It remains to find the asymptotic order of κ_0. By noting that $\frac{d}{dx}\| \mathbf{l}_0(x)\| = \frac{\mathbf{l}_0(x)^T \mathbf{l}_1(x)}{\| \mathbf{l}_0(x)\|}$, it follows that

$$\begin{aligned} \kappa_0 &= \int_a^b \left\| \frac{d}{dx} \frac{\mathbf{l}_0(x)}{\| \mathbf{l}_0(x)\|} \right\| dx \\ &= \int_a^b \sqrt{\frac{\| \mathbf{l}_0(x)\|^2 \| \mathbf{l}_1(x)\|^2 - \{\mathbf{l}_0(x)^T \mathbf{l}_1(x)\}^2}{\| \mathbf{l}_0(x)\|^4}}\, dx. \end{aligned}$$

Next, we need to find the asymptotic orders of the two terms in the equation above. We have that (Fan and Gijbels, 1996, Theorem 3.1)

$$\begin{aligned} \frac{\| \mathbf{l}_1(x)\|^2}{\| \mathbf{l}_0(x)\|^2} &= \frac{1}{h^2} \frac{\varepsilon_2^T S^{-1} S^\star S^{-1} \varepsilon_2}{\varepsilon_1^T S^{-1} S^\star S^{-1} \varepsilon_1} \{1 + o_p(1)\} \\ &= \frac{1}{h^2} \frac{\int K_{1,p}^{\star 2}(t)\, dt}{\int K_{0,p}^{\star 2}(t)\, dt} \{1 + o_p(1)\}, \end{aligned}$$

where $K_{\nu,p}^\star(t) = \varepsilon_{\nu+1}^T S^{-1}(1, t, \ldots, t^p)^T K(t)$ denotes the equivalent kernel in a point t. Further, we have that

$$\begin{aligned} \mathbf{l}_0(x)^T \mathbf{l}_1(x) &= \sum_{i=1}^n l_0\left(\frac{X_i - x}{h}\right) l_1\left(\frac{X_i - x}{h}\right) \\ &= \frac{1}{nh^2 f(x)} \varepsilon_1^T S^{-1} S^\star S^{-1} \varepsilon_2 \{1 + o_p(1)\} \\ &= \frac{1}{nh^2 f(x)} \int K_{0,p}^\star(t) K_{1,p}^{\star T}(t)\, dt \{1 + o_p(1)\}, \end{aligned}$$

and consequently

$$\frac{\{\mathbf{l}_0(x)^T \mathbf{l}_1(x)\}^2}{\| \mathbf{l}_0(x)\|^4} = \frac{1}{h^2} \frac{\left[\int K_{0,p}^\star(t) K_{1,p}^{\star T}(t)\, dt\right]^2}{\left[\int K_{0,p}^{\star 2}(t)\, dt\right]^2} \{1 + o_p(1)\}.$$

Combining all results yields

$$\begin{aligned} \kappa_0 &= \frac{b-a}{h} \sqrt{\left[\frac{\int K_{1,p}^{\star 2}(t)\, dt}{\int K_{0,p}^{\star 2}(t)\, dt} - \frac{\left[\int K_{0,p}^\star(t) K_{1,p}^{\star T}(t)\, dt\right]^2}{\left[\int K_{0,p}^{\star 2}(t)\, dt\right]^2} \right]} \{1 + o_p(1)\} \\ &= O_p(h^{-1}). \end{aligned}$$

It follows that $\kappa_0 = O_p\big(n^{1/(2p+3)}\big)$ for p odd. Consequently, the width of the confidence interval for local polynomial regression (p odd) based on the volume-of-tube formula has asymptotic order $O_p\big(\sqrt{\log n^{2/(2p+3)}}\, n^{-(p+1)/(2p+3)}\big)$. In case p is even, $\kappa_0 = O_p\big(n^{1/(2p+5)}\big)$ and therefore the width of the interval has asymptotic order $O_p\big(\sqrt{\log n^{2/(2p+5)}}\, n^{-(p+2)/(2p+5)}\big)$.

Proof of Theorem 2.3.3. From (2.7), it follows that $c^\star = \sqrt{\log[\kappa_0{}^2\{1 + O_p\big(1\big)\}]} + \widehat{\delta}$. For the boundary region and under the assumptions stated in Theorem 2.3.3, the asymptotic order of the first term of $c^\star$ does not change and is $O_p\big(\sqrt{\log n^{2/(2p+5)}}\big)$ for p even (see proof of Theorem 2.3.2). However, $\widehat{\delta}$ is no longer $O_p\big(1\big)$ due to boundary effects. Noting that the bias in the boundary (left or right) is of order $O_p\big(\widehat{h}^{p+1}\big)$ for p even, it immediately follows that $\widehat{\delta} = O_p\big(n^{1/(2p+5)}\big)$. Since the first term is no longer the leading order term, the width of the interval at the boundary has asymptotic order $O_p\big(\sqrt{n^{2/(2p+5)}}\, n^{-(p+2)/(2p+5)}\big)$ for p even.

2.7 Conclusion and Discussion

In this chapter we derived the asymptotic order of the width of uniform confidence intervals (based on the volume-of-tube formula) for local polynomial regression while considering the bandwidth as a random variable. The asymptotic order does not change from p even to its consecutive p odd degree in the interior region. An interesting fact is that the asymptotic order remains the same in the boundary region for p odd. In case p is even, the asymptotic order of the width of the intervals is slower in the boundary region than in the interior. The simulation study confirmed that the bias-corrected confidence intervals attain nominal coverage.

It is also important to note that the volume-of-tube formula relies on the Gaussian distribution assumption for the errors. If the sample size is large and the central limit theorem applies, the volume-of-tube formula still holds for models with non Gaussian additive independent errors. Loader and Sun (1997) empirically found that the volume-of-tube formula still holds without any modification for an error vector with a spherically symmetric

distribution that has a density in case of an unbiased estimator. According to our simulations, the volume-of-tube formula still performed well in case of a biased estimator and in the presence of independent non Gaussian errors with heavy tails and skewness for large enough sample sizes.

In the context of penalized splines, Wiesenfarth et al. (2012) extended the work of Krivobokova et al. (2010) to additive models with heteroscedastic errors. We believe that extensions of our work to handle heteroscedastic and correlated data offer interesting directions for future research.

2.8 References

Bernardo, J. M. and Smith, A. F. (2009). *Bayesian theory*, volume 405. John Wiley & Sons.

Bickel, P. J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *The Annals of Statistics*, pages 1071–1095.

Chaudhuri, P. and Marron, J. S. (1999). Sizer for exploration of structures in curves. *Journal of the American Statistical Association*, 94(447):807–823.

Chen, S. X. and Qin, Y. S. (2000). Empirical likelihood confidence intervals for local linear smoothers. *Biometrika*, 87(4):946–953.

Chen, S. X. and Qin, Y. S. (2002). Confidence intervals based on local linear smoother. *Scandinavian Journal of Statistics*, 29(1):89–99.

Claeskens, G., Van Keilegom, I., et al. (2003). Bootstrap confidence bands for regression curves and their derivatives. *The Annals of Statistics*, 31(6):1852–1884.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836.

De Brabanter, K., Ferrario, P. G., and Györfi, L. (2014). Detecting ineffective features for nonparametric regression. *Regularization, Optimization, Kernels, and Support Vector Machines*, pages 177–194.

De Brabanter, K., Liu, Y., and Hua, C. (2016). Convergence rates for uniform confidence intervals based on local polynomial regression estimators. *Journal of Nonparametric Statistics*, 28(1):31–48.

Devroye, L., Ferrario, P. G., Györfi, L., and Walk, H. (2013). Strong universal consistent estimate of the minimum mean squared error. In *Empirical Inference*, pages 143–160. Springer.

Eubank, R. L. and Speckman, P. L. (1993). Confidence bands in nonparametric regression. *Journal of the American Statistical Association*, 88(424):1287–1301.

Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American statistical Association*, 87(420):998–1004.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics*, pages 196–216.

Fan, J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *The Annals of Statistics*, pages 2008–2036.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 371–394.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, volume 66. CRC Press.

Hall, P. (1991). On convergence rates of suprema. *Probability Theory and Related Fields*, 89(4):447–455.

Hall, P. (1992). On bootstrap confidence intervals in nonparametric regression. *The Annals of Statistics*, pages 695–711.

Hall, P., Kay, J., and Titterinton, D. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, 77(3):521–528.

Hall, P. and Titterington, D. (1988). On confidence bands in nonparametric density estimation and regression. *Journal of Multivariate Analysis*, 27(1):228–254.

Härdle, W. (1989). Asymptotic maximal deviation of m-smoothers. *Journal of Multivariate Analysis*, 29(2):163–179.

Härdle, W., Hall, P., and Marron, J. (1992). Regression smoothing parameters that are not far from their optimum. *Journal of the American Statistical Association*, 87(417):227–233.

Härdle, W. and Marron, J. (1991). Bootstrap simultaneous error bars for nonparametric regression. *The Annals of Statistics*, pages 778–796.

Hotelling, H. (1939). Tubes and spheres in n-spaces, and a class of statistical problems. *American Journal of Mathematics*, 61(2):440–460.

Huang, L.-S. (1995). *On nonparametric estimation and goodness-of-fit*. PhD thesis, University of North Carolina at Chapel Hill.

Knafl, G., Sacks, J., and Ylvisaker, D. (1985). Confidence bands for regression functions. *Journal of the American Statistical Association*, 80(391):683–691.

Krivobokova, T., Kneib, T., and Claeskens, G. (2010). Simultaneous confidence bands for penalized spline estimators. *Journal of the American Statistical Association*, 105(490):852–863.

Loader, C. (1993). Nonparametric regression, confidence bands and bias correction. *Computing Science and Statistics*, pages 131–131.

Loader, C. and Sun, J. (1997). Robustness of tube formula based confidence bands. *Journal of Computational and Graphical Statistics*, 6(2):242–250.

Müller, H. (1985). Empirical bandwidth choice for nonparametric kernel regression by means of pilot estimators. *Statistical Decisions*, 2:193–206.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142.

Naiman, D. Q. (1986). Conservative confidence bands in curvilinear regression. *The Annals of Statistics*, pages 896–906.

Naiman, D. Q. (1990). Volumes of tubular neighborhoods of spherical polyhedra and statistical inference. *The Annals of Statistics*, pages 685–716.

Neumann, M. H. et al. (1995). Automatic bandwidth choice and confidence intervals in nonparametric regression. *The Annals of Statistics*, 23(6):1937–1959.

Neumann, M. H. and Polzehl, J. (1998). Simultaneous bootstrap confidence bands in nonparametric regression. *Journal of Nonparametric Statistics*, 9(4):307–333.

Oehlert, G. W. (1993). Regional trends in sulfate wet deposition. *Journal of the American Statistical Association*, 88(422):390–399.

Ojeda Cabrera, J. L. (2012). locpol: Kernel local polynomial regression. https://cran.r-project.org/web/packages/locpol/index.html.

Rice, S. (1939). The distribution of the maxima of a random curve. *American Journal of Mathematics*, 61(2):409–416.

Ruppert, D. (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *Journal of the American Statistical Association*, 92(439):1049–1062.

Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *The annals of statistics*, pages 1346–1370.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press.

Silverman, B. W. (2018). *Density estimation for statistics and data analysis*. Routledge.

Stone, C. J. (1977). Consistent nonparametric regression. *The annals of statistics*, pages 595–620.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The annals of Statistics*, pages 1348–1360.

Sun, J. (1993). Tail probabilities of the maxima of gaussian random fields. *The Annals of Probability*, pages 34–71.

Sun, J., Loader, C. R., et al. (1994). Simultaneous confidence bands for linear regression and smoothing. *The Annals of Statistics*, 22(3):1328–1345.

Ullah, A. (1985). Specification analysis of econometric models. *Journal of quantitative economics*, 1:187–209.

Wahba, G. (1983). Bayesian" confidence intervals" for the cross-validated smoothing spline. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 133–150.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372.

Weyl, H. (1939). On the volume of tubes. *American Journal of Mathematics*, 61(2):461–472.

Wiesenfarth, M., Krivobokova, T., Klasen, S., and Sperlich, S. (2012). Direct simultaneous inference in additive models and its application to model undernutrition. *Journal of the American Statistical Association*, 107(500):1286–1296.

Xia, Y. (1998). Bias-corrected confidence bands in nonparametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(4):797–811.

Xia, Y. and Li, W. (2002). Asymptotic behavior of bandwidth selected by the cross-validation method for local polynomial fitting. *Journal of multivariate analysis*, 83(2):265–287.

CHAPTER 3. SMOOTHED NONPARAMETRIC DERIVATIVE ESTIMATION USING WEIGHTED DIFFERENCE QUOTIENTS

Part of the work accepted by Neural Information Processing Systems Conference (Liu and De Brabanter, 2018)

Journal version submitted to Journal of Machine Learning Research

Yu Liu, Kris De Brabanter

Abstract

Derivatives play an important role in bandwidth selection methods (e.g., plug-ins), data analysis and bias-corrected confidence intervals. Therefore, obtaining accurate derivative information is crucial. Although many derivative estimation methods exist, the majority require a fixed design assumption. In this paper, we propose an effective and fully data-driven framework to estimate the first and second order derivative in random design. We establish the asymptotic properties of the proposed derivative estimator, and also propose a fast tuning selection method for the tuning parameters. The performance and flexibility of the method is illustrated via an extensive simulation study.

3.1 Introduction

The next section describes previous methods and current state-of-the-art for nonparametric derivative estimation. Also, we summarize the main differences between derivative estimation in the equispaced and random design for our type of estimator and give a brief overview of local polynomial regression.

3.1.1 Previous work and current state-of-the-art

Since the mid sixties nonparametric density and regression estimation have become a popular and well studied area in statistics. These methods have provided researchers with more flexibility to analyze data without relying on parametric assumptions. Although the literature of nonparametric regression estimators is vast, see e.g., Fan and Gijbels (1996), Györfi et al. (2006) and Tsybakov (2008), derivative estimation also plays an important role in different research areas and applications such as exploration of the structure of data (detecting jump discontinuities (Gijbels and Goderniaux, 2005), revealing important features from curve estimation (Chaudhuri and Marron, 1999), analyzing significant trends (Rondonotti et al., 2007)), comparing regression curves (Park and Kang, 2008), bias-corrected confidence intervals (Eubank and Speckman, 1993; Xia, 1998), analyzing human growth data (Müller, 2012; Ramsay and Silverman, 2007) and neural network pruning (Hassibi and Stork, 1993).

Our proposed methodology provides a data-driven way to estimate derivatives nonparametrically without having to estimate the regression model first. This is especially important when the regression function is difficult to estimate. Although a myriad of papers are published regarding derivative estimation in the mid nineties, many open problems still remain. Ramsay (1998) noted that typically one sees derivatives go wild at the extremes, and the higher the order of the derivative, the wilder the behavior. Further, problems arise in the smoothing parameter or bandwidth selection processes where cross-validation (CV) and generalized CV can be poor guides (Härdle, 1990). Based on Rice (1986), Charnigo et al. (2011) proposed a generalized C_p criterion to determine the smoothing/tuning parameters for derivative estimation for the equispaced design case.

Nonparametric derivative estimation methods can be categorized in three groups: local polynomial regression, regression/smoothing splines, and difference quotients (Müller et al., 1987). Due to the tremendous and well established work done in the field of local polynomial smoothing, the research activity regarding to nonparametric regression and derivative

estimation seems to be somewhat stalled. In local polynomial regression, the derivative can be estimated by the coefficient of the q-th order derivative of the local polynomial regression fitted at point x, i.e. the local slope. Theoretical properties are studied in Fan and Gijbels (1996) and Delecroix and Rosa (1996). The bandwidth choice for the derivative estimator (based on a factor rule) is discussed in Fan and Gijbels (1996). Stone (1985) showed that derivative estimation with splines can achieve the optimal L_2 rate of convergence under mild assumptions. Further asymptotic properties are obtained by Zhou and Wolfe (2000) in the random design setting. However, the smoothing parameter selection problem remained unanswered. Wahba and Wang (1990) noticed that this was particularly difficult for smoothing splines since the smoothing parameter depends on the order of the derivative.

Difference quotient based derivative estimators (Müller et al., 1987; Härdle, 1990) produce a noisy data set which can be smoothed by any nonparametric regression estimator. Smoothing turns out to be quite difficult in practice due to difference quotient's large variance which is $O(n^2)$, where n is the sample size. Therefore, the main goal is to significantly reduce the variance at the cost of a slight bias increase. To obtain such a variance reduction, Iserles (2009) proposed a variance-reducing linear combination of k symmetric difference quotients in the field of numerical mathematics where k is considered to be a tuning parameter. Difference quotients are certainly not new (Müller et al., 1987; Charnigo et al., 2011; De Brabanter et al., 2013), but all results were obtained under the equispaced design assumption. Extending these estimators to the random design setting is possible, however they are no longer consistent for derivatives of order higher than two. This is due to the accumulation of errors associated with the design which will cause higher order terms to blow up. Such an effect is not present when considering equispaced design. Wang and Lin (2015) proposed a sequence of approximate linear regression representations in which the derivative is the intercept term. Although their results are very appealing, they rely on rather stringent assumptions on the regression function. These assumptions are relaxed in Dai et al. (2016) where a linear combination of the dependent variables, depending on two

tuning parameters, are used to obtain derivatives. The variance reducing weights are obtained by solving a constraint optimization problem for which the authors derived a closed form solution. They further showed that the symmetric form used in Charnigo et al. (2011) and De Brabanter et al. (2013) reduces the order of estimation bias without increasing the estimation variance in the interior. They proposed an asymmetric estimator for the derivatives at the boundaries. All results from Wang and Lin (2015) and Dai et al. (2016) assume an equispaced design and both authors do not mention the extension to the random design setting.

In this paper we extend the difference quotient based estimator to the random design to estimate the first order derivative and propose a new consistent estimator for second order derivatives. This framework is flexible so it can be used to extend other difference based estimators in fixed design to the random design. An initial idea of this paper is given in the conference paper of Liu and De Brabanter (2018). Since it is not straightforward to propose an asymptotically consistent estimator for the general case, we will first provide a framework to estimate the first and second order derivative in the uniform random design and then generalize it to arbitrary distributions. Because this method produces a new data set containing correlated errors, we use the local polynomial regression estimator with bandwidth selection method of De Brabanter et al. (2018) to smooth the noisy derivatives and derive the asymptotic properties of the smoothed derivative estimators.

The paper is organized as follows. We discuss the main theoretical differences between equispaced and random design for this type of estimator and give a short description of local polynomial regression. Section 2 illustrates the first order derivative estimation based on variance reducing weighted difference quotients. Bias, variance and pointwise consistency are established. In addition, bandwidth selection and behavior at the boundary for noisy derivative estimators are also described. Finally, we discuss how to smooth the data with correlated noise and study the asymptotic properties of the smoothed derivatives. Section 3 extends the framework to second order derivatives. In section 4, we conduct Monte

Carlo experiments to compare the proposed methodology with smoothing splines and local polynomial regression. Finally, Section 5 states the conclusions and future work.

3.1.2 Equispaced design vs. random design

Consider the data (X_1, Y_1),...,(X_n, Y_n) which form an independent and identically distributed (i.i.d.) sample from a population (X, Y), where $X_i \in \mathcal{X} = [a, b] \subseteq \mathbb{R}$ and $Y \in \mathbb{R}$ for all $i = 1, \ldots, n$. In the equispaced design case, the response variables are assumed to satisfy

$$Y_i = m(x_i) + e_i, \quad i = 1, \ldots, n, \tag{3.1}$$

where $x_1, \ldots, x_n$ are nonrandom numbers and $x_{i+1} - x_i = (b - a)/(n - 1)$ is constant for all i. In this setting, the regression function is given by $m(x) = \mathbf{E}[Y]$ and we assume that $\mathbf{E}[e] = 0$ and $\mathbf{Var}[e] = \sigma_e^2 < \infty$. In contrast to the equispaced design, the X are random variables in random design and are generated from an unknown distribution F. Consider the following model

$$Y_i = m(X_i) + e_i, \quad i = 1, \ldots, n, \tag{3.2}$$

where the regression function is given by $m(X) = \mathbf{E}[Y|X = x]$ and assume that $\mathbf{E}[e] = 0$, $\mathbf{Var}[e] = \sigma_e^2 < \infty$, X and e are independent. The derivative estimators discussed in Charnigo et al. (2011) and De Brabanter et al. (2013) use the symmetric property $x_{i+j} - x_i = x_i - x_{i-j}$ since they both assumed equispaced design. However, in the random design this property no longer holds which introduces extra estimation error. In addition, it is fairly complicated to obtain an asymptotic expression for the difference $X_{i+j} - X_i$ when the X's are generated from an unknown distribution, leading to theoretical difficulties in obtaining asymptotic properties of the derivative estimator.

3.1.3 Local Polynomial Regression

The local polynomial regression estimator in an arbitrary point x is given by minimizing the following weighted least squares problem (Fan and Gijbels, 1996)

$$\min_{\beta_j \in \mathbb{R}} \sum_{i=1}^{n} \Big\{ Y_i - \sum_{j=0}^{p} \beta_j \, (X_i - x)^j \Big\}^2 K_h(X_i - x), \tag{3.3}$$

where β_j are the solutions to the weighted least squares problem, K is a symmetric probability density function with $K_h(\cdot) = K(\cdot/h)/h$. Note that $\hat{m}^{(q)}(x) = q!\hat{\beta}_q$ is an estimator for the q-th order derivative $m^{(q)}(x)$, $q = 0, 1, \ldots, p$. In matrix notation the solution is

$$\hat{\beta} = (\mathbf{X}^T \, \mathbf{W} \, \mathbf{X})^{-1} \, \mathbf{X}^T \, \mathbf{W} \, \mathbf{Y},$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\beta = (\beta_0, \ldots, \beta_p)^T$ and

$$\mathbf{X} = \begin{pmatrix} 1 & (X_1 - x) & \cdots & (X_1 - x)^p \\ \vdots & \vdots & & \vdots \\ 1 & (X_n - x) & \cdots & (X_n - x)^p \end{pmatrix},$$

with $\mathbf{W}$ a $n \times n$ diagonal matrix of weights based on the kernel function and the bandwidth h

$$\mathbf{W} = \text{diag}\{K_h(X_i - x)\}.$$

3.2 First Order Derivative Estimation

Müller et al. (1987) introduced the first order difference quotients to produce noisy derivative data. However, all their results are obtained for the equispaced design setting. In case of random design, their estimator for the first order (noisy) derivative at design point X_i is denoted by $\hat{q}_i^{(1)}$ and is

$$\hat{q}^{(1)}(X_i) = \hat{q}_i^{(1)} = \frac{Y_i - Y_{i-1}}{X_i - X_{i-1}}. \tag{3.4}$$

Although very appealing (quasi unbiased) and intuitive, this estimator has major drawbacks, i.e. (i) a large variance and (ii) difficulties in studying its asymptotic properties in random

design. The variance is $O(n^2)$ and $O_p\{(X_i - X_{i-1})^{-2}\}$ for the equispaced and random design respectively. In the latter case, it is obvious that this can be very large when the distance between two neighboring X is small. Consequently, reducing variance in these type of estimators is paramount and can be accomplished by means of a variance-reducing linear combination of symmetric difference quotients. Second, In order to discuss the asymptotic properties of this different quotient, we need to obtain an asymptotic expression for the difference $X_i - X_{i-1}$ which is not trivial in the random design setting. However, in a special case, i.e. $X = U \sim \mathcal{U}(0,1)$ and arranging the random variables in order of magnitude according to U (order statistics), the asymptotic properties of the first order quotient (3.4) can be obtained. In what follows, $\mathcal{U}(0,1)$ denotes the uniform distribution between 0 and 1. For the sake of simplicity, we will first discuss a special case, i.e. $U = X \sim \mathcal{U}(0,1)$, before we formulate the estimator for arbitrary distributions.

3.2.1 Approach based on order statistics

Consider n bivariate data forming an i.i.d sample from a population (U, Y) and further assume $U \sim \mathcal{U}(0,1)$. Arrange the bivariate data (U, Y) in order of magnitude according to U, i.e. $U_{(1)} < U_{(2)} < \ldots < U_{(n)}$ where $U_{(i)}$, $i = 1, \ldots, n$ is the i-th order statistic. In order to avoid ties and hence division by zero we also require $U_{(i)} \neq U_{(j)}$ for $i \neq j$. The first order difference quotient (3.4) is

$$\hat{q}^{(1)}(U_{(i)}) = \hat{q}_i^{(1)} = \frac{Y_i - Y_{i-1}}{U_{(i)} - U_{(i-1)}}. \tag{3.5}$$

Next, consider the model

$$Y_i = r(U_{(i)}) + e_i, \tag{3.6}$$

where $r(U) = \mathbf{E}[Y|U = u]$ is the regression function and assume $\mathbf{E}[e] = 0$, $\mathbf{Var}[e] = \sigma_e^2 < \infty$, U and e are independent. Now the difference $U_{(i)} - U_{(i-1)}$ is the difference of uniform order statistics and it is well-known that (David and Nagaraja, 2003, p. 14)

$$U_{(s)} - U_{(r)} \sim \text{Beta}(s - r, n - s + r + 1) \quad \text{for } s > r.$$

This result immediately leads to lemma 1.

Lemma 1 *Let* $U \overset{i.i.d.}{\sim} \mathcal{U}(0,1)$. *Arrange the random variables in order of magnitude* $U_{(1)} < U_{(2)} < \cdots < U_{(n)}$. *Then, for* $i > j$

$$U_{(i+j)} - U_{(i-j)} = \frac{2j}{n+1} + O_p\left(\sqrt{\frac{j}{n^2}}\right),$$

$$U_{(i+j)} - U_{(i)} = \frac{j}{n+1} + O_p\left(\sqrt{\frac{j}{n^2}}\right)$$

and

$$U_{(i)} - U_{(i-j)} = \frac{j}{n+1} + O_p\left(\sqrt{\frac{j}{n^2}}\right).$$

Proof: see 3.6.1 ■

Assume r is twice continuously differentiable on $[0,1]$. A Taylor expansion of $r(U_{(i\pm j)})$ in a neighborhood of $U_{(i)}$ gives

$$r(U_{(i\pm j)}) \quad = \quad r(U_{(i)}) + r^{(1)}(U_{(i)})(U_{(i\pm j)} - U_{(i)}) + O_p\left(\frac{j^2}{n^2}\right). \tag{3.7}$$

Using Lemma 1 for $j = 1$ and (3.7)

$$\mathbf{E}[\hat{q}_i^{(1)} | U_{(i-1)}, U_{(i)}] \quad = \quad \mathbf{E}\left[\frac{Y_i - Y_{i-1}}{U_{(i)} - U_{(i-1)}} | U_{(i-1)}, U_{(i)}\right] = r^{(1)}(U_{(i)}) + o_p(1)$$

for $n \to \infty$ and

$$\mathbf{Var}\left[\hat{q}_i^{(1)} | U_{(i-1)}, U_{(i)}\right] = \mathbf{Var}\left[\frac{Y_i - Y_{i-1}}{U_{(i)} - U_{(i-1)}} | U_{(i-1)}, U_{(i)}\right] = \frac{2\sigma_e^2}{(U_{(i)} - U_{(i-1)})^2} = O_p(n^2).$$

It is immediately clear that this estimator is asymptotically unbiased. However, the variance of this estimator can be arbitrary large and hence it will be difficult to estimate the smoothed derivative function. A possible way to reduce the variance is described in Iserles (2009) and used in Charnigo et al. (2011) and De Brabanter et al. (2013) which involves a combination of symmetric difference quotients around the i-th point. Our proposed derivative estimator for random design involving uniform order statistics is

$$\hat{Y}_i^{(1)} = \sum_{j=1}^{k} w_{i,j} \cdot \left(\frac{Y_{i+j} - Y_{i-j}}{U_{(i+j)} - U_{(i-j)}}\right), \tag{3.8}$$

where the weights $w_{i,1}, \ldots, w_{i,k}$ sum up to one. Note that (3.8) is valid for $k+1 \leq i \leq n-k$ and hence $k \leq (n-1)/2$. For the boundary regions, i.e. $2 \leq i \leq k$ and $n-k+1 \leq i \leq n-1$, the estimator (3.8) needs to be modified and is discussed in Section 3.2.5. The estimator (3.8) does not provide results for $\hat{Y}_1^{(1)}$ and $\hat{Y}_n^{(1)}$. One can ignore these two points from consideration or have them coincide with $\hat{Y}_2^{(1)}$ and $\hat{Y}_{n-1}^{(1)}$ (see Charnigo et al. (2011)).

The following proposition states the optimal weights $w_{i,j}$, optimal in the sense of minimizing the variance of the estimator (3.8).

Proposition 1 *For $k+1 \leq i \leq n-k$ and under model (3.6), the weights $w_{i,j}$ that minimize the variance of (3.8), satisfying $\sum_{j=1}^{k} w_{i,j} = 1$, are given by*

$$w_{i,j} = \frac{(U_{(i+j)} - U_{(i-j)})^2}{\sum_{l=1}^{k} (U_{(i+l)} - U_{(i-l)})^2}, \quad j = 1, \ldots, k. \tag{3.9}$$

Proof: see 3.6.2 ■

For fixed i, the j-th weight (3.9) is proportional to the inverse variance of the difference quotient $\frac{Y_{i+j} - Y_{i-j}}{U_{(i+j)} - U_{(i-j)}}$ in (3.8). At first sight, these weights seem to be different than the weights obtained by Charnigo et al. (2011) and De Brabanter et al. (2013) for the equispaced design case. Plugging in the difference $u_{i+j} - u_{i-j} = 2j(b-a)/(n-1)$ for equispaced design on $[a, b]$ yields

$$w_{i,j} = \frac{(u_{i+j} - u_{i-j})^2}{\sum_{l=1}^{k} (u_{i+l} - u_{i-l})^2} = \frac{\frac{4j^2}{(n-1)^2}}{\frac{4}{(n-1)^2} \sum_{l=1}^{k} l^2} = \frac{6j^2}{k(k+1)(2k+1)}.$$

These are exactly the weights obtained in Charnigo et al. (2011) & De Brabanter et al. (2013). This shows that the weights for equispaced design are a special case of the weights in Proposition 1. However, one parameter still remains unknown, i.e. k, the number of symmetric difference quotients (around i). Theorem 1 (asymptotic conditional bias and variance) provides valuable insights how to choose k.

3.2.2 Asymptotic properties of the first order derivative estimator

The following theorems establish the asymptotic conditional bias and variance of our proposed estimator (3.8) for the interior points, i.e. $k+1 \leq i \leq n-k$. In what follows we denote $\mathbb{U} = (U_{(i-j)}, \ldots, U_{(i+j)})$ for $i > j\,, i+j \leq n$ and $j = 1, \ldots, k$.

Theorem 1 *Under model (3.6) and assume r is twice continuously differentiable on $[0,1]$ and $k \to \infty$ as $n \to \infty$. Then, for uniform random design on $[0,1]$ and the weights in Proposition 1, the conditional (absolute) bias and conditional variance of (3.8) are*

$$\left|\text{bias}\big[\hat{Y}_i^{(1)}|\mathbb{U}\big]\right| \leq \sup_{u\in[0,1]} |r^{(2)}(u)| \frac{3k(k+1)}{4(n+1)(2k+1)} + o_p(n^{-1}k)$$

and

$$\mathbf{Var}\big[\hat{Y}_i^{(1)}|\mathbb{U}\big] = \frac{3\sigma_e^2(n+1)^2}{k(k+1)(2k+1)} + o_p(n^2k^{-3})$$

uniformly for $k+1 \leq i \leq n-k$.

Proof: see 3.6.3 ■

From Theorem 1, the pointwise consistency of (3.8) immediately follows.

Corollary 1 *Under the assumptions of Theorem 1, $k \to \infty$ as $n \to \infty$ such that $n^{-1}k \to 0$ and $n^2k^{-3} \to 0$. Then, for $\sigma_e^2 < \infty$ and the weights given in Proposition 1, we have for any $\varepsilon > 0$*

$$\mathbf{P}(|\hat{Y}_i^{(1)} - r^{(1)}(U_{(i)})| \geq \varepsilon) \to 0$$

for $k+1 \leq i \leq n-k$.

Proof: see 3.6.4 ■

According to Theorem 1 and Corollary 1, the conditional bias and conditional variance of (3.8) tend to zero and $k \to \infty$ faster than $O(n^{2/3})$, but slower than $O(n)$. Next, we develop a rule-of-thumb tuning method for k such that $k = O(n^{4/5})$ and the fastest possible

rate at which $\mathbf{E}[(\hat{Y}_i^{(1)} - r^{(1)}(U_{(i)}))^2|\mathbb{U}] \to 0$ (L_2 rate of convergence) is $O_p(n^{-2/5})$. Using Jensen's inequality, similar results can be shown for the L_1 rate of convergence, i.e.

$$\mathbf{E}\big[\big|\hat{Y}_i^{(1)} - r^{(1)}(U_{(i)})\big| \mid \mathbb{U}\big] \leq \Big|\text{bias}\big[\hat{Y}_i^{(1)}|\mathbb{U}\big]\Big| + \sqrt{\mathbf{Var}\big[\hat{Y}_i^{(1)}|\mathbb{U}\big]} = O_p(n^{-1/5}).$$

3.2.3 Tuning k

Parameter k in (3.8) controls the bias-variance trade-off. Based on Theorem 1, we choose k that minimizes the asymptotic upper bound of the conditional mean integrated squared error (MISE). The result is given in Corollary 2.

Corollary 2 *Under the assumptions of Theorem 1 and denote* $\mathcal{B} = \sup_{u\in[0,1]} |r^{(2)}(u)|$, *then* k *that minimizes the asymptotic upper bound of the conditional MISE is*

$$k_{\text{opt}} \quad = \quad \underset{k\in\mathbb{N}^+\setminus\{0\}}{\arg\min} \left\{ \mathcal{B}^2 \frac{9k^2(k+1)^2}{16(n+1)^2(2k+1)^2} + \frac{3\sigma_e^2(n+1)^2}{k(k+1)(2k+1)} \right\} = O(n^{4/5}).$$

Proof: see 3.6.5 ■

Corollary 2 provides a fast and easy parameter tuning method in practice, however some unknown quantities still need to be estimated. The error variance can be estimated by Hall's $\sqrt{n}$-consistent estimator (Hall et al., 1990)

$$\hat{\sigma}_e^2 = \frac{1}{n-2}\sum_{i=1}^{n-2}(0.809Y_i - 0.5Y_{i+1} - 0.309Y_{i+2})^2.$$

The second unknown quantity $\mathcal{B}$ can be (roughly) estimated with a local polynomial regression estimator of order $p = 3$. The performance of our proposed model is not so sensitive to the accuracy of $\mathcal{B}$, thus a rough estimate of the second order derivative is sufficient. By plugging in these two estimators for σ_e^2 and $\mathcal{B}$ in Corollary 2, the optimal value k_{opt} can be obtained for example by a grid search over the integer set $[1, \lfloor\frac{n-1}{2}\rfloor]$ where $\lfloor x\rfloor$ denotes the greatest integer less than or equal to x..

3.2.4 Asymptotic order of the conditional bias and continuous differentiability of r

In Theorem 1, we bounded the conditional bias above. From a theoretical point of view, it is helpful to derive an exact expression for the conditional bias and discuss its dependence on the continuous differentiability of the true regression function r. It also allows us to compare with the bias in fixed design and explain the extra bias due to the asymmetric differences $U_{(i+j)} - U_{(i)} \neq U_{(i)} - U_{(i-j)}$ in random design. Assume the first $q+1$ derivatives of r exist on $[0,1]$. A Taylor series of $r(U_{(i\pm j)})$ in a neighborhood of U_i and using Lemma 1 yields

$$\begin{aligned} r(U_{(i+j)}) &= r(U_{(i)}) + \sum_{l=1}^{q} \frac{1}{l!}(U_{(i+j)} - U_{(i)})^l r^{(l)}(U_{(i)}) + O_p(U_{(i+j)} - U_{(i)})^{q+1} \\ &= r(U_{(i)}) + \sum_{l=1}^{q} \frac{1}{l!}(U_{(i+j)} - U_{(i)})^l r^{(l)}(U_{(i)}) + O_p\{(j/n)^{q+1}\} \\ &= r(U_{(i)}) + \sum_{l=1}^{q} \frac{1}{l!}(U_{(i-j)} - U_{(i)})^l r^{(l)}(U_{(i)}) + O_p\{(j/n)^{q+1}\}. \end{aligned}$$

and

$$\begin{aligned} r(U_{(i-j)}) &= r(U_{(i)}) + \sum_{l=1}^{q} \frac{1}{l!}(U_{(i-j)} - U_{(i)})^l r^{(l)}(U_{(i)}) + O_p(U_{(i-j)} - U_{(i)})^{q+1} \\ &= r(U_{(i)}) + \sum_{l=1}^{q} \frac{1}{l!}(U_{(i-j)} - U_{(i)})^l r^{(l)}(U_{(i)}) + O_p\{(j/n)^{q+1}\}. \end{aligned}$$

Using Lemma 1, assume $k \to \infty$ as $n \to \infty$, and for the weights in Proposition 1 we obtain the asymptotic order of the exact conditional bias for different values of q.

$$\text{bias}\left[\hat{Y}_i^{(1)} | \mathbb{U}\right] = \begin{cases} O_p\left(\frac{k}{n}\right) & , \quad q = 1. \\ \\ O_p\left(\max\left\{\frac{k^{\frac{1}{2}}}{n}, \frac{k^2}{n^2}\right\}\right), & q \geq 2 \end{cases}$$

The proof is given in 3.6.6 For $q = 1$ (i.e. r is twice continuously differentiable), the leading order of exact conditional bias is the same as that of the bias upperbound given in Theorem 1. For $q = 2$, r is three times continuously differentiable on $[0,1]$, the exact bias

achieves smaller order than $O_p(k/n)$. Unfortunately, adding additional assumptions on the differentiability of r, i.e. $q > 2$, will no longer improve the asymptotic rate of the bias. This can be seen as follows: for $q \geq 2$, the bias is

$$\text{bias}\big[\hat{Y}_i^{(1)}|\mathbb{U}\big] = \frac{\sum_{j=1}^{k}(U_{(i+j)} - U_{(i-j)})\big[\sum_{l=2}^{q} \frac{r^{(l)}(U_{(i)})\{(U_{(i+j)}-U_{(i)})^l - (U_{(i-j)}-U_{(i)})^l\}}{l!} + O_p\big\{(j/n)^{q+1}\big\}\big]}{\sum_{p=1}^{k}(U_{(i+p)} - U_{(i-p)})^2}.$$

This can be split into two terms: odd and even with $l \geq 2$

$$\text{bias}_{\text{odd}}[\hat{Y}_i^{(1)}|\mathbb{U}\big] = O_p\left(\frac{k^2}{n^2}\right) \qquad \text{and} \qquad \text{bias}_{\text{even}}[\hat{Y}_i^{(1)}|\mathbb{U}\big] = O_p\left(\frac{k^{\frac{1}{2}}}{n}\right).$$

resulting in

$$\begin{aligned} \text{bias}\big[\hat{Y}_i^{(1)}|\mathbb{U}\big] &= \text{bias}_{\text{odd}}[\hat{Y}_i^{(1)}|\mathbb{U}\big] + \text{bias}_{\text{even}}[\hat{Y}_i^{(1)}|\mathbb{U}\big] \\ &= O_p\left\{\max\left(\frac{k^2}{n^2}, \frac{k^{\frac{1}{2}}}{n}\right)\right\}. \end{aligned}$$

In fixed design, $\text{bias}_{\text{even}} = 0$ due to symmetry: $u_{(i+j)} - u_{(i)} = u_{(i)} - u_{(i-j)}$. Unfortunately, in the random design, we cannot remove $\text{bias}_{\text{even}}$. It is this fact that will lead to the inconsistency of third and higher order derivatives if these estimators are defined in a fully recursive way as in Charnigo et al. (2011).

3.2.5 Boundary correction

We discussed the proposed estimator at the interior points and in this section we provide a simple but effective boundary correction. Points with index $i < k + 1$ and $i > n - k$ are points located at the left and right boundary respectively. Since there are not enough k pairs of neighbors at the boundary, we use a weighted linear combination of $k(i)$ pairs of points U_i instead, where $k(i) = i - 1$ for the left boundary and $k(i) = n - i$ for the right boundary. This is the approach of Charnigo et al. (2011) and De Brabanter et al. (2013). The first order derivative estimator at the boundary is obtained by replacing k with $k(i)$ in (3.8) and weights in Proposition 1. From Section 3.2.4, we know that if r is three times continuously differentiable on $[0, 1]$ the asymptotic order of the conditional bias at the boundary is $O_p\big\{\max\big(\frac{k(i)^2}{n^2}, \frac{k(i)^{1/2}}{n}\big)\big\}$, which is smaller than for the interior points.

However, the asymptotic order of the conditional variance is $O_p\left\{\frac{3\sigma_e^2(n+1)^2}{k(i)(k(i)+1)(2k(i)+1)}\right\}$ and attains $O_p(n^2)$, as i is close to either 2 or $n-1$.

In order to reduce the variance at the boundary we propose the following modification to (3.8). For points at the left boundary, $i < k+1$, consider the estimator

$$\hat{Y}_i^{(1)} = \sum_{j=1}^{k(i)} w_{i,j} \cdot \left(\frac{Y_{i+j} - Y_{i-j}}{U_{(i+j)} - U_{(i-j)}}\right) + \sum_{j=k(i)+1}^{k} w_{i,j} \cdot \left(\frac{Y_{i+j} - Y_i}{U_{(i+j)} - U_{(i)}}\right) \tag{3.10}$$

with

$$w_{i,j} = \begin{cases} \dfrac{(U_{(i+j)} - U_{(i-j)})^2}{\sum_{l=1}^{k(i)}(U_{(i+l)} - U_{(i-l)})^2 + \sum_{l=k(i)+1}^{k}(U_{(i+l)} - U_{(i)})^2}, & 1 \le j \le k(i); \\ \dfrac{(U_{(i+j)} - U_{(i)})^2}{\sum_{l=1}^{k(i)}(U_{(i+l)} - U_{(i-l)})^2 + \sum_{l=k(i)+1}^{k}(U_{(i+l)} - U_{(i)})^2}, & k(i) < j \le k. \end{cases}$$

This modification leads to

$$\text{bias}[\hat{Y}_i^{(1)}|\mathbb{U}] = O_p\left\{\max\left(\frac{k(i)^{7/2}}{k^3 n}, \frac{k(i)^5}{k^3 n^2}, \frac{k - k(i)}{n}\right)\right\}$$

and

$$\mathbf{Var}[\hat{Y}_i^{(1)}|\mathbb{U}] = O_p\left\{\max\left(\frac{n^2}{k^3}, \frac{n^2(k-k(i))^2}{k^4}\right)\right\}.$$

The proof is given in 3.6.7 The $\text{bias}[\hat{Y}_i^{(1)}|\mathbb{U}] \to 0$ when $n \to \infty$ indicating that (3.10) is still asymptotically unbiased at the boundary. Worst case scenario, the variance is of the order $O_p(n^2/k^2)$ which is smaller than $O_p(n^2)$. A similar result can be obtained for the right boundary.

3.2.6 Smoothing the noisy derivatives

Noisy first order derivative estimators (3.8) and (3.10) have two problems: (i) derivative estimators contain the noise coming from the unknown errors $e_i, i = 1, \ldots, n$ in model (3.6) and (ii) derivative estimators can only be evaluated at the design points $U_{(i)}, i = 1, \ldots, n$. Hence some type of smoothing will be needed to remove the noise and evaluate the derivative

in an arbitrary point. The first order derivative estimator (3.8) can be written as

$$\hat{Y}_i^{(1)} = \sum_{j=1}^{k} w_{i,j} \cdot \left(\frac{r(U_{(i+j)}) - r(U_{(i-j)})}{U_{(i+j)} - U_{(i-j)}} \right) + \sum_{j=1}^{k} w_{i,j} \cdot \left(\frac{e_{i+j} - e_{i-j}}{U_{(i+j)} - U_{(i-j)}} \right) \tag{3.11}$$

where the second term $\tilde{e}_i = \sum_{j=1}^{k} w_{i,j} \cdot \left(\frac{e_{i+j} - e_{i-j}}{U_{(i+j)} - U_{(i-j)}} \right)$ is the new error. It is clear that $\tilde{e}_i, i = 2, \ldots, n-1$ are correlated and for the generated derivatives $\hat{Y}_i^{(1)}, i = 2, \ldots, n-1$, the i.i.d. assumption of the errors is no longer valid. Hence, bandwidth selection for any nonparametric smoothing method becomes increasingly difficult (Opsomer et al., 2001; De Brabanter et al., 2018). In this paper we use the idea of De Brabanter et al. (2018) by using a kernel K such that $K(0) = 0$. By using such a kernel, De Brabanter et al. (2018) have shown that under mild assumptions, the effect of the correlation on the bandwidth selection process is removed without any prior knowledge about the correlation structure.

For interior points $k+1 \leq i \leq n-k$, all $\hat{Y}_i^{(1)}$ are asymptotic consistent estimators. Without loss of generality, we show the properties of the smoothed derivative estimator in the interior. The local polynomial estimator at an arbitrary point u_0 is

$$\hat{r}^{(1)}(u_0) = \boldsymbol{\epsilon}_1^T \hat{\beta} = \boldsymbol{\epsilon}_1^T \, \mathbf{S}_n^{-1} \, \mathbf{U}_{\mathrm{u}}{}^T \, \mathbf{W}_{\mathrm{u}} \, \hat{\mathbf{Y}}^{(1)} \tag{3.12}$$

where $\boldsymbol{\epsilon}_1 = (1, 0, \ldots, 0)^T$ is a unit vector with 1 in the first position and $\hat{r}^{(q+1)}(u_0) = q! \hat{\beta}_q$. $\hat{\mathbf{Y}}^{(1)} = (\hat{Y}_{k+1}^{(1)}, \ldots, \hat{Y}_{n-k}^{(1)})$, $\mathbf{W}_{\mathrm{u}}$ is the diagonal matrix of weights, i.e. $\mathrm{diag}\{K_h(U_{(i)} - u_0)\}$ with kernel K, bandwidth h and $K_h(\cdot) = K(\cdot/h)/h$, $\mathbf{S}_n = \mathbf{U}_{\mathrm{u}}{}^T \, \mathbf{W}_{\mathrm{u}} \, \mathbf{U}_{\mathrm{u}}$, and

$$\mathbf{U}_{\mathrm{u}} = \begin{pmatrix} 1 & (U_{(k+1)} - u_0) & \cdots & (U_{(k+1)} - u_0)^p \\ \vdots & \vdots & & \vdots \\ 1 & (U_{(n-k)} - u_0) & \cdots & (U_{(n-k)} - u_0)^p \end{pmatrix}.$$

The term $\tilde{e}_i = \sum_{j=1}^{k} w_{i,j} \cdot \left(\frac{e_{i+j} - e_{i-j}}{U_{(i+j)} - U_{(i-j)}} \right)$ in (3.11) satisfies $\mathbf{E}[\tilde{e}_i | U] = 0$ and $\mathbf{Cov}(\tilde{e}_i, \tilde{e}_j | U_{(i)}, U_{(j)}) = \sigma_{\tilde{e}}^2 \rho_n(U_{(i)} - U_{(j)})$ for $i \neq j$ with $\sigma_{\tilde{e}}^2 < \infty$ and ρ_n is a stationary correlation function satisfying $\rho_n(0) = 1, \rho_n(u) = \rho_n(-u)$ and $|\rho_n(u)| \leq 1$ for all u. The subscript n allows the correlation function ρ_n to shrink as $n \to \infty$ (De Brabanter et al., 2018). In what follows, we denote $\widetilde{\mathbb{U}} = (U_{(1)}, \ldots, U_{(n)})$. Under the following assumptions:

- Assumption 1. $h \to 0$ and $nh \to \infty$ as $n \to \infty$;
- Assumption 2. there exists a constant $K_{\max}$ such that $|K(x)| < K_{\max}$, and $K(x) \geq 0$ for all x;
- Assumption 3. K is symmetric and Lipschitz continuous at 0;
- Assumption 4. $\lim_{|u|\to\infty} |u|^l K(u) < \infty$ for $l = 0, \ldots, p$;
- Assumption 5. the correlation function ρ_n is an element of a sequence $\{\rho_n\}$ with the following properties for all n: there exists constants ρ_{max} and ρ_c such that $n \int |\rho_n(x)|dx < \rho_{max}$ and $\lim_{n\to\infty} n \int \rho_n(x)dx = \rho_c$; and for any sequence $\epsilon_n > 0$ satisfying $n\epsilon_n \to \infty$,

$$n \int_{|x|\geq\epsilon_n} |\rho_n(x)|dx \to 0, n \to \infty,$$

we derive the conditional bias and variance of the smoothed derivative $\hat{r}^{(1)}(u_0)$ by applying Theorem 1 in De Brabanter et al. (2018).

Theorem 2 *Assume $f(u) > 0$ and let $f(\cdot)$ and $r^{(p+1)}(\cdot), p \geq 1$ be continuous in a neighbourhood of u_0. Under the Assumptions $1-5$, $k \to \infty$ as $n \to \infty$, $\sigma_e^2 < \infty$ and the weights given in Proposition 1, the conditional bias and conditional variance of (3.12) for p odd is*

$$\begin{aligned}
\operatorname{bias}\left[\hat{r}^{(1)}(u_0)|\widetilde{\mathbb{U}}\right] &\leq \boldsymbol{\epsilon}_1^T \mathbf{S}^{-1} \left[\frac{c_p}{(p+1)!} r^{(p+2)}(u_0) h^{p+1} + \mathcal{B} \frac{3k(k+1)}{4(n+1)(2k+1)} \tilde{c}_p \right] \{1 + o_p(1)\} \\
&= \left[\left(\int t^{p+1} K_0^*(t)dt \right) \frac{1}{(p+1)!} r^{(p+2)}(u_0) h^{p+1} \right. \\
&+ \left. \mathcal{B} \frac{3k(k+1)}{4(n+1)(2k+1)} \left(\int K_0^*(t)dt \right) \right] \{1 + o_p(1)\}
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{Var}\left[\hat{r}^{(1)}(u_0)|\widetilde{\mathbb{U}}\right] &= \frac{3\sigma_e^2(n+1)^2}{k(k+1)(2k+1)} \frac{1 + f(u_0)\rho_c}{h(n-2k)f(u_0)} \boldsymbol{\epsilon}_1^T \mathbf{S}^{-1} \mathbf{S}^* \mathbf{S}^{-1} \boldsymbol{\epsilon}_1 \{1 + o_p(1)\} \\
&= \int K_0^{*2}(t)dt \frac{3\sigma_e^2(n+1)^2}{k(k+1)(2k+1)} \frac{1 + f(u_0)\rho_c}{h(n-2k)f(u_0)} \{1 + o_p(1)\}
\end{aligned}$$

where $\mathcal{B} = \sup_{u\in[0,1]} |r^{(2)}(u)|$, $\mathbf{S} = (\mu_{i+j})_{0\leq i,j\leq p}$ with $\mu_j = \int u^j K(u)du$, $\mathbf{S}^* = (\nu_{i+j})_{0\leq i,j\leq p}$ with $\nu_j = \int u^j K^2(u)du$, $c_p = (\mu_{p+1}, \ldots, \mu_{2p+1})^T$, $\tilde{c}_p = (\mu_0, \mu_1, \ldots, \mu_p)^T$, $\boldsymbol{\epsilon}_1 = (1, 0, \ldots, 0)^T$,

and the equivalent kernel $K_0^*(t) = \boldsymbol{\epsilon}_1^T S^{-1}(1, t, \ldots, t^p)^T K(t)$.

Proof: see 3.6.8 ■

The asymptotic upper bound of the conditional MISE is minimized for $h = O(n^{-\frac{2}{5p+6}})$ and $k = O(n^{\frac{3p+4}{5p+6}})$. The corresponding L_2 rate of convergence is $O_p(n^{-\frac{4p+4}{5p+6}})$. In this paper, we will not use the variance-bias trade-off in Theorem 2 to select the bandwidth h and the parameter k simultaneously, since it requires estimating ρ_c, which is not straightforward. To have an easy and efficient tuning method at the cost of a slower rate of convergence, we use Corollary 2 to select k then select bandwidth h as follows. First, use kernel $\overline{K}(u) = (2/\sqrt{\pi})u^2 \exp(-u^2)$ to obtain the bandwidth h_b by minimizing the residual sum of squares (RSS) of interior points $(U_{(i)}, \hat{Y}_i^{(1)})$ with $k+1 \leq i \leq n-k$, i.e. $\text{RSS}(h_b) = (n-2k)^{-1} \sum_{i=k+1}^{n-k} (\hat{r}^{(1)}(U_{(i)}) - \hat{Y}_i^{(1)})^2$. RSS does not contain the boundary points, since noisy derivatives $\hat{Y}_i^{(1)}$ at the boundary have larger variance. Second, as bimodal kernels introduce extra error in the estimation due to their non-optimality we overcome this issue by using $\hat{h}_b$ as a pilot bandwidth and relate it to a bandwidth $\hat{h}$ of a more optimal (unimodal) kernel, say the Gaussian kernel. As shown in De Brabanter et al. (2018), this can be achieved without any extra smoothing step. For local cubic regression, the relation between the bimodal and unimodal bandwidth is

$$\hat{h} = 1.01431 \hat{h}_b$$

when using $\overline{K}(u) = (2/\sqrt{\pi})u^2 \exp(-u^2)$ and $K(u) = (1/\sqrt{2\pi}) \exp(-u^2/2)$ as bimodal and unimodal kernel respectively.

From Theorem 2, the pointwise consistency of (3.12) for p odd immediately follows.

Corollary 3 *Under the assumptions of Theorem 2,* $h \to 0$ *and* $nh \to \infty$ *as* $n \to \infty$, $k \to \infty$ *as* $n \to \infty$ *such that* $n^{-1}k \to 0$ *and* $nk^{-3}h^{-1} \to 0$. *Then, for* $\sigma_e^2 < \infty$ *and the weights given in Proposition 1, we have for any* $\varepsilon > 0$

$$\mathbf{P}(|\hat{r}^{(1)}(u_0) - r^{(1)}(u_0)| \geq \varepsilon) \to 0$$

Proof: see 3.6.9 ■

For $k = O(n^{4/5})$ in Corollary 2 and $h = O(n^{-1/(2p+3)})$ for p odd, then by Corollary 1 in De Brabanter et al. (2018), we have

$$\mathbf{E}[(\hat{r}^{(1)}(u_0) - r^{(1)}(u_0))^2 | \widetilde{\mathbb{U}}] = O_p(n^{-2/5})$$

Using Jensen's inequality, the L_1 rate of convergence is $O_p(n^{-1/5})$.

3.2.7 Generalizing results for first order derivatives to arbitrary distributions

It is possible to find a closed form expression for the distribution of the differences $X_{(i+j)} - X_{(i-j)}$ with $X \overset{i.i.d}{\sim} F$ where F is unknown and continuous (David and Nagaraja, 2003) such that the density function $f(x) = F'(x)$. Since this result is quite unattractive from a theoretical point of view, we advocate the use of the probability integral transform (PIT) (Casella and Berger, 2002)

$$F(X) \sim U(0, 1). \tag{3.13}$$

By using the probability integral transform we know that the new data set $(F(X_{(1)}), Y_1), \ldots, (F(X_{(n)}), Y_n)$ has the same distribution as $(U_{(1)}, Y_1), \ldots, (U_{(n)}, Y_n)$. This leads to the original setting of uniform order statistics discussed earlier. The final step is to transform back to the original space. In order for this step to work, we need the existence of a density f. Since $m(X) = r(F(X))$ and by the chain rule

$$\frac{dm(X)}{dX} = \frac{dr(U)}{dU}\frac{dU}{dX} = f(X)\frac{dr(U)}{dU}, \tag{3.14}$$

yielding $m^{(1)}(X) = f(X)r^{(1)}(U)$ which is the smoothed version of the first order derivative in the original space. In practice, the distribution F and density f need to be estimated giving $\widehat{m}^{(1)}(X) = \widehat{f}(X)\widehat{r}^{(1)}(U)$. In this paper we use the kernel density estimator (Rosenblatt, 1956; Parzen, 1962) to estimate the density f and distribution F with plug-in bandwidth (**?**).

3.3 Higher Order Derivatives

In practice, first and second order derivatives are widely used. However, higher order derivatives become progressively more difficult to estimate, i.e. they suffer from higher bias and variance and consequently slower rate of convergence. In this section, we construct an efficient estimator for the second order derivative and discuss its asymptotic properties. A similar procedure can be applied to estimate derivatives with order higher than two.

3.3.1 Asymptotic results for second order noisy derivatives under standard uniform distribution

As before, assume $U \sim \mathcal{U}(0,1)$, and (U,Y) are sorted according to ascending order of U. We define the second order noisy derivative estimator as

$$\hat{Y}_i^{(2)} = 2\sum_{j=1}^{k_2} w_{i,j,2} \frac{\left(\frac{Y_{i+j+k_1}-Y_{i+j}}{U_{(i+j+k_1)}-U_{(i+j)}} - \frac{Y_{i-j-k_1}-Y_{i-j}}{U_{(i-j-k_1)}-U_{(i-j)}}\right)}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}} \tag{3.15}$$

where both k_1 and k_2 are positive integers and the weights $w_{i,j,2}$ sum up to one. The subscript 2 is used to indicate the weight for the second order derivative. Let ${}^{+}\hat{Y}_{i+j}^{(1)} = \frac{Y_{i+j+k_1}-Y_{i+j}}{U_{(i+j+k_1)}-U_{(i+j)}}$, which is an asymptotically conditional unbiased estimator of $r^{(1)}(U_{(i+j)})$. $\mathbf{Var}[{}^{+}\hat{Y}_{i+j}^{(1)}|\widetilde{\mathbb{U}}] = O_p(\frac{\sigma_e^2 n^2}{k_1^2})$, where the parameter k_1 controls the variance of ${}^{+}\hat{Y}_{i+j}^{(1)}$. The left superscript "+" indicates the estimator only uses data on the right hand side of Y_{i+j}. Similarly ${}^{-}\hat{Y}_{i-j}^{(1)} = \frac{Y_{i-j-k_1}-Y_{i-j}}{U_{(i-j-k_1)}-U_{(i-j)}}$ is an asymptotically conditional unbiased estimator of $r^{(1)}(U_{(i-j)})$ and the conditional variance is $O_p(\frac{\sigma_e^2 n^2}{k_1^2})$. The left superscript "−" indicates the estimator only uses data on the left hand side of Y_{i-j}. Hence the estimator (3.15) can be rewritten as

$$\hat{Y}_i^{(2)} = \sum_{j=1}^{k_2} w_{i,j,2} \frac{{}^{+}\hat{Y}_{i+j}^{(1)} - {}^{-}\hat{Y}_{i-j}^{(1)}}{C_{i,j,k_1}}$$

where $C_{i,j,k_1} = (U_{(i+j+k_1)} + U_{(i+j)} - U_{(i-j-k_1)} - U_{(i-j)})/2$. By defining the estimator in this way, the variance of $\hat{Y}_i^{(2)}$ is reduced by decreasing the correlation between different

quotients. Assume r is three times continuously differentiable on the compact interval $[0,1]$, $k_1 \to \infty$ and $k_2 \to \infty$ as $n \to \infty$. Then appying Lemma 1 gives

$$\begin{aligned}
\mathbf{E}[{}^{+}\hat{Y}_{i+j}^{(1)} - {}^{-}\hat{Y}_{i-j}^{(1)} | \widetilde{\mathbb{U}}] &= \frac{r(U_{(i+j+k_1)}) - r(U_{(i+j)})}{U_{(i+j+k_1)} - U_{(i+j)}} - \frac{r(U_{(i-j-k_1)}) - r(U_{(i-j)})}{U_{(i-j-k_1)} - U_{(i-j)}} \\
&= r^{(1)}(U_{(i+j)}) + \frac{1}{2} r^{(2)}(U_{(i+j)})(U_{(i+j+k_1)} - U_{(i+j)})\{1 + o_p(1)\} \\
&- r^{(1)}(U_{(i-j)}) - \frac{1}{2} r^{(2)}(U_{(i-j)})(U_{(i-j-k_1)} - U_{(i-j)})\{1 + o_p(1)\} \\
&= \frac{1}{2} r^{(2)}(U_{(i)})(U_{(i+j+k_1)} + U_{(i+j)} - U_{(i-j-k_1)} - U_{(i-j)})\{1 + o_p(1)\}
\end{aligned}$$

where C_{i,j,k_1} is chosen such that each individual quotient $\frac{{}^{+}\hat{Y}_{i+j}^{(1)} - {}^{-}\hat{Y}_{i-j}^{(1)}}{C_{i,j,k_1}}$, $j = 1, \ldots, k_2$ is an unbiased estimator of the second order derivative $r^{(2)}(U_{(i)})$.

The exact weight is selected to be proportional to the inverse of the conditional variance of each quotient $\frac{{}^{+}\hat{Y}_{i+j}^{(1)} - {}^{-}\hat{Y}_{i-j}^{(1)}}{C_{i,j,k_1}}$

$$\tilde{w}_{i,j,2} = \frac{1/\mathbf{Var}\left[\frac{\left(\frac{Y_{i+j+k_1} - Y_{i+j}}{U_{(i+j+k_1)} - U_{(i+j)}} - \frac{Y_{i-j-k_1} - Y_{i-j}}{U_{(i-j-k_1)} - U_{(i-j)}}\right)}{U_{(i+j+k_1)} + U_{(i+j)} - U_{(i-j-k_1)} - U_{(i-j)}} | \widetilde{\mathbb{U}}\right]}{\sum_{j=1}^{k_2} 1/\mathbf{Var}\left[\frac{\left(\frac{Y_{i+j+k_1} - Y_{i+j}}{U_{(i+j+k_1)} - U_{(i+j)}} - \frac{Y_{i-j-k_1} - Y_{i-j}}{U_{(i-j-k_1)} - U_{(i-j)}}\right)}{U_{(i+j+k_1)} + U_{(i+j)} - U_{(i-j-k_1)} - U_{(i-j)}} | \widetilde{\mathbb{U}}\right]}$$

By Lemma 1, the leading order of the weight $\tilde{w}_{i,j,2}$ is

$$w_{i,j,2} = \frac{(2j + k_1)^2}{\sum_{j=1}^{k_2} (2j + k_1)^2} \tag{3.16}$$

such that $\tilde{w}_{i,j,2} = w_{i,j,2}\{1 + o_p(1)\}$ for $k_1 \to \infty$ and $k_2 \to \infty$ as $n \to \infty$. Similar to the first order noisy derivative, boundary issues arise in (3.15) when $i < \sum_{j=1}^{2} k_j + 1$ or $i > n - \sum_{j=1}^{2} k_j$. Theorem 3 states the asymptotic conditional bias and variance of (3.15) using the weights (3.16). It is difficult to get the exact asymptotic expression for the conditional bias and variance of the noisy second order derivative estimator. Therefore, we provide a suitable upperbound.

Theorem 3 *Under model (3.6) and assume r is three times continuously differentiable on $[0,1]$, $k_1 \to \infty$ and $k_2 \to \infty$ as $n \to \infty$. Then, for the weights (3.16), the conditional*

(absolute) bias and the conditional variance of (3.15) *are bounded above*

$$\left|\text{bias}\left[\hat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\right]\right| \leq \frac{\sup_{u\in[0,1]}|r^{(3)}(u)|}{n+1}\frac{2\sum_{j=1}^{k_2}j^3+3k_1\sum_{j=1}^{k_2}j^2+\frac{5}{3}k_1^2\sum_{j=1}^{k_2}j+\frac{1}{3}k_1^3k_2}{4\sum_{j=1}^{k_2}j^2+k_1^2k_2+4k_1\sum_{j=1}^{k_2}j}\{1+o_p(1)\}$$

and

$$\mathbf{Var}[\hat{Y}_i^{(2)}|\widetilde{\mathbb{U}}] \leq \frac{4(n+1)^4\sigma_e^2}{k_1^2\sum_{j=1}^{k_2}(2j+k_1)^2}\{1+o_p(1)\}$$

uniformly for $\sum_{j=1}^{2}k_j+1 \leq i \leq n-\sum_{j=1}^{2}k_j$.

Proof: see 3.6.10

From Theorem 3 the pointwise consistency easily follows

Corollary 4 *Under the assumptions of Theorem 3 and for the weight sequence defined in* (3.16), $k_1 \to \infty$ *and* $k_2 \to \infty$ *as* $n \to \infty$ *such that* $n^{-1}k_1 \to 0$, $n^{-1}k_2 \to 0$, $n^4k_1^{-2}k_2^{-3} \to 0$ *and* $n^4k_1^{-4}k_2^{-1} \to 0$, *it follows for any* $\epsilon > 0$

$$\mathbf{P}(\left|\hat{Y}_i^{(2)} - r^{(2)}(U_{(i)})\right| > \epsilon) \to 0.$$

Proof: see 3.6.11

Assuming the order of k_1 is the same as the order of k_2, then according to Theorem 3 and Corollary 4, the conditional bias and conditional variance of (3.15) tends to zero as $k_1 \to \infty$ and $k_2 \to \infty$ faster than $O_p(n^{4/5})$ but slower than $O_p(n)$. It is easy to show that the fastest possible rate at which $\mathbf{E}[(\hat{Y}_i^{(2)} - r^{(2)}(U_i))^2| \mid \widetilde{\mathbb{U}}] \to 0$ (L_2 rate of convergence) is $O_p(n^{-2/7})$ and the fastest rate is attained for $k_1 = O(n^{6/7})$ and $k_2 = O(n^{6/7})$. Using Jensen's inequality, similar results can be shown for the L_1 rate of convergence, i.e.

$$\mathbf{E}\left[\left|\hat{Y}_i^{(2)} - r^{(2)}(U_{(i)}\right| \mid \widetilde{\mathbb{U}}\right] \leq \left|\text{bias}\left[\hat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\right]\right| + \sqrt{\mathbf{Var}\left[\hat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\right]} = O_p(n^{-1/7}).$$

3.3.2 Tuning k_1 and k_2

As for the first order derivative with one tuning parameter, the second order derivative has two, which control the bias-variance trade-off. Based on the asymptotic upperbounds of the bias and variance in Theorem 3, we choose k_1 and k_2 as follows.

Corollary 5 *Under the assumptions of Theorem 3 and denote* $\mathcal{B}_2 = \sup_{u\in[0,1]} |r^{(3)}(u)|$, *then* k_1 *and* k_2 *that minimize the asymptotic upper bound of the conditional MISE are*

$$(k_1,k_2)_{\text{opt}} = \underset{k_1,k_2\in\mathbb{N}^+\setminus\{0\}}{\arg\min} \left\{ \left(\frac{\mathcal{B}_2}{n+1} \frac{2\sum_{j=1}^{k_2} j^3 + 3k_1\sum_{j=1}^{k_2} j^2 + \frac{5}{3}k_1^2\sum_{j=1}^{k_2} j + \frac{1}{3}k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1\sum_{j=1}^{k_2} j} \right)^2 + \frac{4(n+1)^4\sigma_e^2}{k_1^2\sum_{j=1}^{k_2}(2j+k_1)^2} \right\}$$

Proof: see 3.6.12 ■

The second unknown quantity $\mathcal{B}_2$ can be (roughly) estimated with a local polynomial regression estimator of order $p = 4$. By plugging in two estimators for σ_e^2 (Hall et al., 1990) and $\mathcal{B}_2$ in Corollary 5, the optimal value pair $(k_1,k_2)_{\text{opt}}$ can be obtained using a grid search (or any other optimization method) over a Cartesian product set.

3.3.3 Exact bias

Although we use the upper bound of the absolute conditional bias to tune the parameters in the estimator of second order derivatives, the exact conditional bias for noisy second order derivative estimation (3.15) is also important. It provides a clear comparision with the second order derivative estimator used in fixed design (Charnigo et al., 2011; De Brabanter et al., 2013) and illustrates why we can not use a similar framework in random design.

Adapting the fixed design framework from Charnigo et al. (2011) and De Brabanter et al. (2013) to random design under the standard uniform distribution for the q-th order derivative

$$\hat{Y}_i^{(q)} = \sum_{i=1}^{k_q} w_{i,j} \frac{\hat{Y}_{i+j}^{(q-1)} - \hat{Y}_{i-j}^{(q-1)}}{U_{i+j} - U_{i-j}}, \quad q = 1,2,\ldots \tag{3.17}$$

where $k_1, k_2, \ldots, k_q$ are tuning parameters. Due to the asymmetry $U_{(i+j)} - U_{(i)} \neq U_{(i)} - U_{(i-j)}$ in random design, extra bias will be introduced in the first order noisy derivative estimator $\hat{Y}_i^{(1)}, i = 2, \ldots, n-1$. Using the recursive relation in (3.17), the extra bias will accumulate as q increases in random design. The estimator (3.17) is no longer a consistent estimator when $q > 2$. The exact bias of the proposed second order derivative estimator in (3.15) is smaller and is given by

$$\text{bias}\big[\hat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\big] = O_p\left(\max\left\{\frac{k_1^{\frac{1}{2}}}{n}, \frac{k_2^{\frac{1}{2}}}{n}, \frac{k_1^2}{n^2}, \frac{k_2^2}{n^2}\right\}\right) \tag{3.18}$$

The proof is given in 3.6.13 The boundary issue still arises for the second order derivative estimator. Similar to Section 3.2.5, k_1 and k_2 are no longer constants for points at the boundary $i < 1 + k_1 + k_2$ and $i > n - k_1 - k_2$.

3.3.4 Smoothing the noisy second order derivatives

The second order derivative estimator (3.15) can be written as

$$\begin{aligned}\hat{Y}_i^{(2)} &= 2\sum_{j=1}^{k_2} w_{i,j,2} \frac{\left(\frac{r(U_{(i+j+k_1)}) - r(U_{(i+j)})}{U_{(i+j+k_1)} - U_{(i+j)}} - \frac{r(U_{(i-j-k_1)}) - r(U_{(i-j)})}{U_{(i-j-k_1)} - U_{(i-j)}}\right)}{U_{(i+j+k_1)} + U_{(i+j)} - U_{(i-j-k_1)} - U_{(i-j)}} \\ &+ 2\sum_{j=1}^{k_2} w_{i,j,2} \frac{\left(\frac{e_{i+j+k_1} - e_{i+j}}{U_{(i+j+k_1)} - U_{(i+j)}} - \frac{e_{i-j-k_1} - e_{i-j}}{U_{(i-j-k_1)} - U_{(i-j)}}\right)}{U_{(i+j+k_1)} + U_{(i+j)} - U_{(i-j-k_1)} - U_{(i-j)}}\end{aligned} \tag{3.19}$$

where the second term is the new error term and is denoted as $\acute{e}_i$. It is clear that for $\acute{e}_i, i = 3, \ldots, n-2$ the i.i.d. assumption is no longer valid. Similar to Section 3.2.6, we apply a kernel K such that $K(0) = 0$ to remove the effects of correlation on the bandwidth selection process (De Brabanter et al., 2018).

Without loss of the generality, we show the properties of the smoothed estimator for the interior points $\sum_{j=1}^{2} k_j + 1 \leq i \leq n - \sum_{j=1}^{2} k_j$. For an arbitrary point u_0

$$\hat{r}^{(2)}(u_0) = \boldsymbol{\epsilon}_1^T \hat{\beta} = \boldsymbol{\epsilon}_1^T \, \mathbf{S}_n^{-1} \, \mathbf{U}_{\text{u}}{}^T \, \mathbf{W}_{\text{u}} \, \hat{\mathbf{Y}}^{(2)} \tag{3.20}$$

where $\boldsymbol{\epsilon}_1 = (1, 0, \ldots, 0)^T$ is a unit vector with 1 in the first position. $\hat{\mathbf{Y}}^{(2)} = (\hat{Y}^{(2)}_{k_1+k_2+1}, \ldots, \hat{Y}^{(2)}_{n-k_1-k_2})$, $\mathbf{W}_{\text{u}}$ is the diagonal matrix of weights, i.e. $\text{diag}\{K_h(U_{(i)} - u_0)\}$ with kernel K, bandwidth

h and $K_h(\cdot) = K(\cdot/h)/h$, $\mathbf{S}_n = \mathbf{U_u}^T \mathbf{W_u} \mathbf{U_u}$, and

$$\mathbf{U_u} = \begin{pmatrix} 1 & (U_{(k_1+k_2+1)} - u_0) & \cdots & (U_{(k_1+k_2+1)} - u_0)^p \\ \vdots & \vdots & & \vdots \\ 1 & (U_{(n-k_1-k_2)} - u_0) & \cdots & (U_{(n-k_1-k_2)} - u_0)^p \end{pmatrix}$$

with $\acute{e}_i$ in (3.19) satisfying $\mathbf{E}[\acute{e}_i|U] = 0$ and $\mathbf{Cov}(\acute{e}_i, \acute{e}_j|U_{(i)}, U_{(j)}) = \sigma^2_{\acute{e}} \rho'_n(U_{(i)} - U_{(j)})$ for $i \neq j$ with $\sigma^2_{\acute{e}} < \infty$ and ρ'_n is a stationary correlation function satisfying $\rho'_n(0) = 1, \rho'_n(u) = \rho'_n(-u)$ and $|\rho'_n(u)| \leq 1$ for all u. Applying Theorem 1 in De Brabanter et al. (2018) yields the following theorem.

Theorem 4 *Assume $f(u) > 0$ and let $f(\cdot)$ and $r^{(p+1)}(\cdot), p \geq 2$ be continuous in a neighbourhood of u_0. Under the Assumptions 1−5 and $k_1 \to \infty$, $k_2 \to \infty$ as $n \to \infty$. For $\sigma^2_e < \infty$ and the weights given in (3.16), the conditional bias and conditional variance of (3.20) for p odd are bounded above*

$$\begin{aligned} \text{bias}\big[\hat{r}^{(2)}(u_0)|\widetilde{\mathbb{U}}\big] &\leq \boldsymbol{\epsilon}_1^T \mathbf{S}^{-1} \left[\frac{\mathcal{B}_2}{n+1} \frac{2\sum_{j=1}^{k_2} j^3 + 3k_1 \sum_{j=1}^{k_2} j^2 + \frac{5}{3}k_1^2 \sum_{j=1}^{k_2} j + \frac{1}{3}k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1 \sum_{j=1}^{k_2} j} \tilde{c}_p \right. \\ &+ \left. \frac{c_p}{(p+1)!} r^{(p+3)}(u_0) h^{p+1} \right] \{1 + o_p(1)\} \\ &= \left[\left(\int K_0^*(t)dt \right) \frac{\mathcal{B}_2}{n+1} \frac{2\sum_{j=1}^{k_2} j^3 + 3k_1 \sum_{j=1}^{k_2} j^2 + \frac{5}{3}k_1^2 \sum_{j=1}^{k_2} j + \frac{1}{3}k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1 \sum_{j=1}^{k_2} j} \right. \\ &+ \left. \left(\int t^{p+1} K_0^*(t)dt \right) \frac{1}{(p+1)!} r^{(p+3)}(u_0) h^{p+1} \right] \{1 + o_p(1)\} \end{aligned}$$

and

$$\begin{aligned} \mathbf{Var}\big[\hat{r}^{(2)}(u_0)|\widetilde{\mathbb{U}}\big] &\leq \frac{4(n+1)^4 \sigma_e^2}{k_1^2 \sum_{j=1}^{k_2} (2j+k_1)^2} \frac{1 + f(u_0)\rho'_c}{h(n-2k_1-2k_2)f(u_0)} \boldsymbol{\epsilon}_1^T \mathbf{S}^{-1} \mathbf{S}^* \mathbf{S}^{-1} \boldsymbol{\epsilon}_1 \{1 + o_p(1)\} \\ &= \frac{4(n+1)^4 \sigma_e^2}{k_1^2 \sum_{j=1}^{k_2} (2j+k_1)^2} \frac{1 + f(u_0)\rho'_c}{h(n-2k_1-2k_2)f(u_0)} \left(\int K_0^{*2}(t)dt \right) \{1 + o_p(1)\} \end{aligned}$$

where $\mathcal{B}_2 = \sup_{u \in [0,1]} |r^{(3)}(u)|$.

Proof: see 3.6.14 ■

If the order of k_1 is the same as the order of k_2, the asymptotic upper bound of the conditional MISE is minimized at $h = O(n^{-\frac{2}{7p+8}})$, $k_1 = O(n^{\frac{5p+6}{7p+8}})$ and $k_2 = O(n^{\frac{5p+6}{7p+8}})$ and L_2 rates of convergence is $O_p(n^{-\frac{4p+4}{7p+8}})$. The way to select the bandwidth h is the same as for the first order smoothed derivative estimator. We use Corollary 5 to select k_1 and k_2, and then select bandwidth h by minimizing the RSS in order to avoid estimating ρ_c'. The proposed estimator with a two step parameter tuning is still asymptotic consistent. From Theorem 4, the pointwise consistency of (3.20) for p odd immediately follows.

Corollary 6 *Under the assumptions of Theorem 4, $h \to 0$ and $nh \to \infty$ as $n \to \infty$, $k_1 \to \infty$ and $k_2 \to \infty$ as $n \to \infty$ such that $n^{-1}k_1 \to 0$, $n^{-1}k_2 \to 0$, $n^3 k_1^{-2} k_2^{-3} h^{-1} \to 0$ and $n^3 k_1^{-4} k_2^{-1} h^{-1} \to 0$. Then, for $\sigma_e^2 < \infty$ and the weights given in (3.16), we have for any $\varepsilon > 0$*

$$\mathbf{P}(|\hat{r}^{(2)}(u_0) - r^{(2)}(u_0)| \geq \varepsilon) \to 0$$

Proof: Analogous to the proof in 3.6.9 ■.

Assume k_1 and k_2 have the same order, then for $k_1 = O(n^{6/7})$ and $k_2 = O(n^{6/7})$ in Corollary 5, and $h = O(-n^{1/(2p+3)})$ for p odd from Corollary 1 in De Brabanter et al. (2018), the L_2 rates of convergence is

$$\mathbf{E}[(\hat{r}^{(2)}(u_0) - r^{(2)}(u_0))^2 | \widetilde{\mathbb{U}}] = O_p(n^{-2/7})$$

Using Jensen's inequality, the L_1 rate of convergence is $O_p(n^{-1/7})$.

3.3.5 Generalizing noisy second order derivative to arbitrary distributions

As before, we use the Probability Integral Transform (PIT) as in (3.13) to transform the random variables X to U. Assume the second order derivative of $F(X)$ exists, taking the derivative on both sides of $m(X) = r(F(X))$ with respect to X

$$\frac{d^2 m}{dX^2} = \frac{d}{dX}\left(\frac{dr}{dU}\frac{dU}{dX}\right) = \frac{d}{dX}\left(f(X) r^{(1)}(U)\right) = f^{(1)}(X) r^{(1)}(U) + f(X) r^{(2)}(U) \qquad (3.21)$$

leading to $m^{(2)}(X) = f^{(1)}(X)r^{(1)}(U)+f(X)r^{(2)}(U)$, where $f^{(1)}(X) = \frac{df(X)}{dX}$. In practice, the distribution F and density f need to be estimated. In this paper we use the kernel density estimator (Rosenblatt, 1956; Parzen, 1962) to estimate the density f and distribution F.

For higher order derivatives $Y^{(q)}$, noisy estimates can be constructed in a similar way.

3.4 Simulation Study

In Theorem 2 and Theorem 4, $\hat{r}^{(1)}(\cdot)$ and $\hat{r}^{(2)}(\cdot)$ are based on noisy derivative data for interior points. In the simulation, we include the noisy derivative data at the boundary to obtain the local polynomial regression estimator for the final smoothed derivatives.

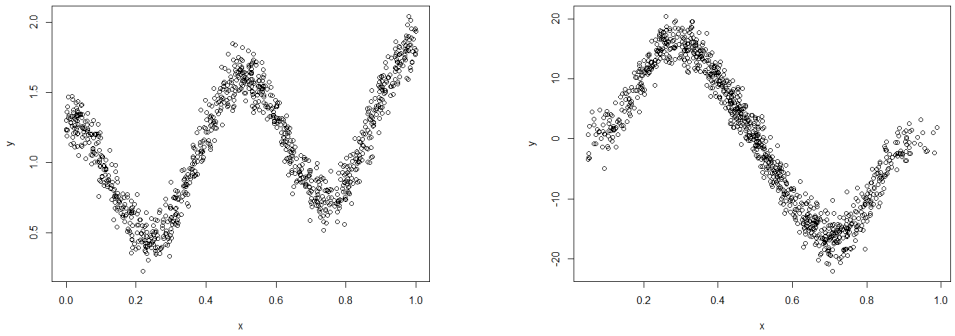
3.4.1 First order derivative estimation

Consider the following two functions

$$m(X) = \cos^2(2\pi X) + \log(4/3 + X) \quad \text{for} \quad X \sim \mathcal{U}(0, 1) \tag{3.22}$$

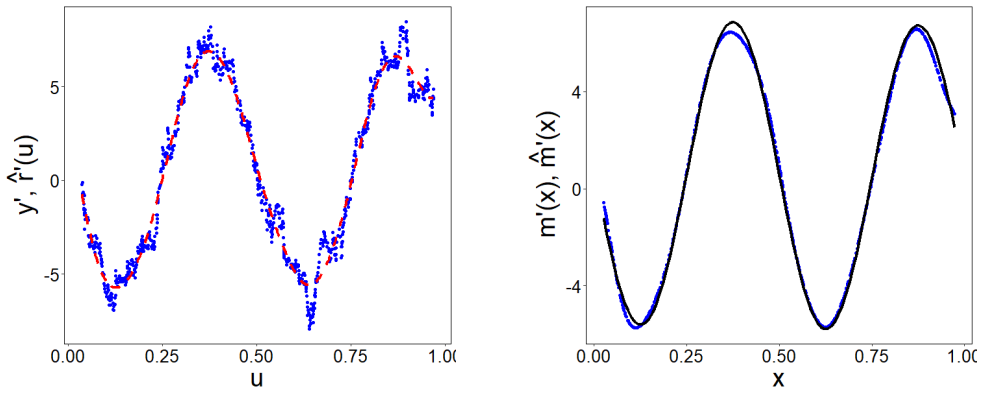
$$m(X) = 50e^{-8(1-2X)^4}(1 - 2X) \quad \text{for} \quad X \sim \text{beta}(2, 2). \tag{3.23}$$

In all simulations, we estimate the density f and distribution F using kernel methods (R package `ks` (Duong, 2018)). The tuning parameter k is selected based on Corollary 5 over a positive integer set $\{1, 2, \ldots, 499\}$. We use local cubic regression ($p = 3$) with bimodal kernel to initially smooth the data. Bandwidths h were selected from the set $\{0.04, 0.045, \ldots, 0.1\}$ for both (3.22) and (3.23) and corrected for a unimodal Gaussian kernel. The sample size for both models is $n = 1000$ with $e \sim N(0, 0.1^2)$ and $e \sim N(0, 2^2)$ for (3.22) and (3.23) respectively. Figure 3.1 shows the raw data (X, Y) for both model (3.22) and model (3.23). Figure 3.2 and 3.3 show the first order noisy derivative (blue dots), the true first order derivative (full line) and smoothed first order derivative (dashed line) for both model (3.22) and model (3.23) separately.



(a) model (3.22) (b) model (3.23)

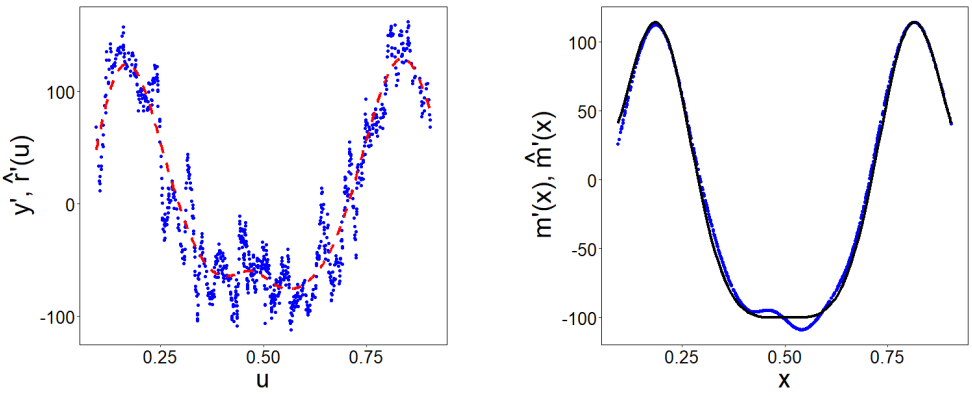
Figure 3.1: Raw Data



(a) Illustration of the transformation $U = \hat{F}(X)$

(b) Back transform according to $\widehat{m}^{(1)}(X) = \widehat{f}(X)\widehat{r}^{(1)}(U)$

Figure 3.2: First order noisy derivative (dots) of model (3.22) based on $k = 26$, smoothed derivative based on local cubic regression (dashed line) and true derivative (full line). (a) First order noisy derivative (dots) and smoothed derivative of $r(U)$ (dashed line), after probability integral transform of the original data. (b) True first order derivative (full line) and the proposed smoothed derivative of $m(X)$ (dashed line) in the original space. Boundary points are not shown for visual purposes.



(a) Illustration of the transformation $U = \hat{F}(X)$

(b) Back transform according to $\widehat{m}^{(1)}(X) = \widehat{f}(X)\widehat{r}^{(1)}(U)$

Figure 3.3: First order noisy derivative (dots) of model (3.23) based on $k = 22$, smoothed derivative based on local cubic regression (dashed line) and true derivative (full line). (a) First order noisy derivative (dots) and smoothed derivative of $r(U)$ (dashed line), after probability integral transform of the original data. (b) True first order derivative (full line) and the proposed smoothed derivative of $m(X)$ (dashed line) in the original space. Boundary points are not shown for visual purposes.

Next, we compare the proposed methodology with several popular methods for non-parametric derivative estimation, i.e. the local slope of the local polynomial regression with $p = 2$, $p = 3$ (R package `locpol` (Ojeda Cabrera, 2012)) and penalized smoothing splines (R package `pspline` (Ramsey and Ripley, 2017)). The order of the local polynomial is recommended to be $p = 2$ since p minus the order of the derivative is odd (Fan and Gijbels, 1996). In case of penalized smoothing splines, cubic splines were used. For the Monte Carlo study, we constructed data sets of size $n = 700$ and generated the function

$$m(X) = \sqrt{X(1-X)}\sin((2.1\pi)/(X+0.05)) \quad \text{for} \quad X \sim \mathcal{U}(0.25, 1)$$

100 times according to model (3.2) with $e \sim N(0, 0.2^2)$. Bandwidths were selected from the set $\{0.03, 0.035, \ldots, 0.07\}$ and corrected for a unimodal Gaussian kernel. In order to remove the effect of boundary issues on the performance for all three methods, we use the

adjusted mean absolute error as a performance measure defined as

$$\text{MAEadjusted} = \frac{1}{650}\sum_{i=26}^{675} |\hat{m}_n'(X_i) - m'(X_i)|.$$

The first three boxplots in Figure 3.4 represent the performance of the proposed estimator, local quadratic polynomial regression and penalized smoothing splines. The proposed model has a similar performance as the local polynomial regression with p=2 (locpol2) and cubic penalized smoothing splines (psplines). To illustrate the loss of accuracy due to estimation of the density f and distribution F we use the true density and distribution to compute the derivative in the fourth boxplot.

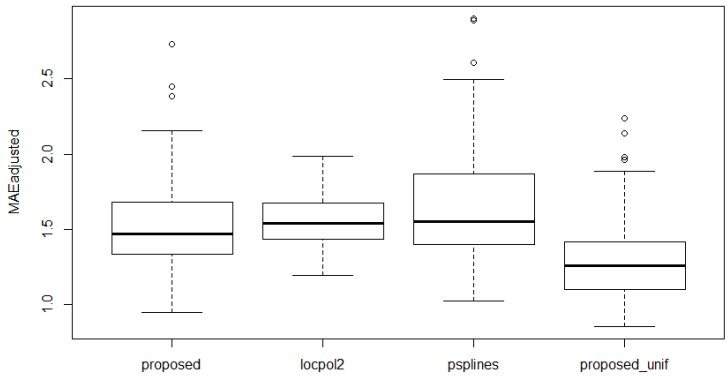


Figure 3.4: Result of the Monte Carlo study for the proposed methodology, local polynomial regression and penalized smoothing splines for first order derivative estimation.

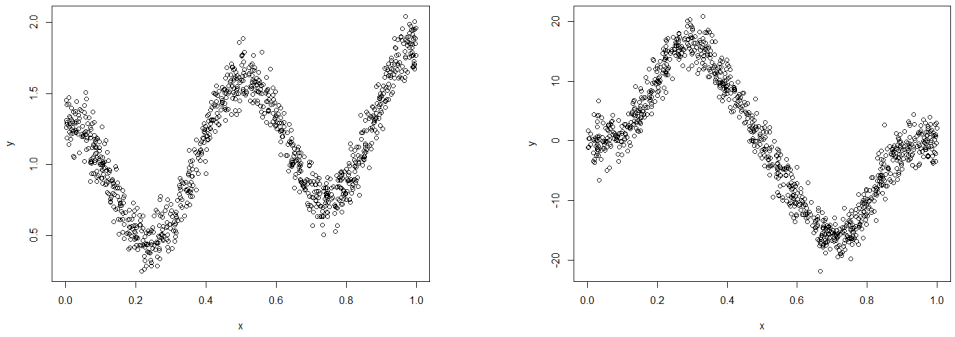
3.4.2 Second order derivative estimation

Similar to the first order derivative, the tuning parameters k_1 and k_2 could be determined by minimizing the criterion in corollary 5 through grid search over a product set. We use local cubic regression ($p = 3$) with a kernel K such that $K(0) = 0$ to smooth the noisy second order derivatives. The bandwidth obtained with the kernel K such that $K(0) = 0$ is then corrected for a unimodal kernel. The second order derivative estimation for any distribution is given in (3.21). In the simulation, we only show the performance of the

proposed second order derivative estimator under the assumption that $X \sim \mathcal{U}[0,1]$. For model (3.23), we change the assumption on the distribution of X as follows

$$m(X) = 50e^{-8(1-2X)^4}(1-2X) \quad \text{for} \qquad X \sim \mathcal{U}(0,1). \tag{3.24}$$

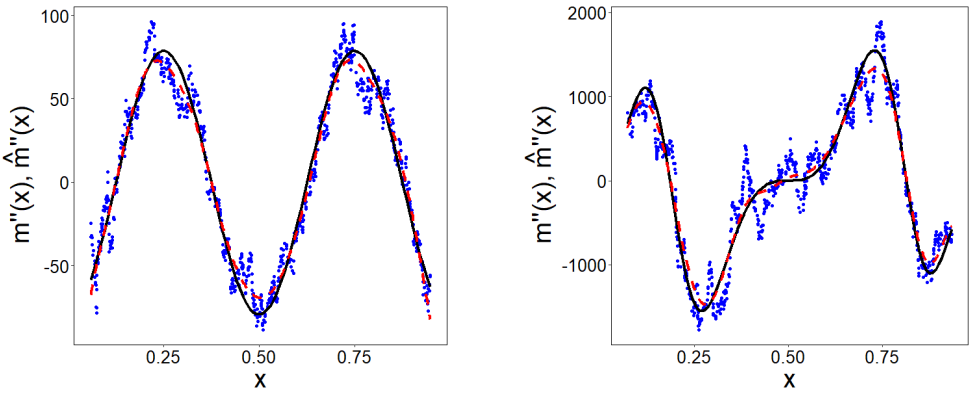
Figure 3.5 shows the raw data (X, Y) for both models (3.22) and (3.24). The sample size is taken to be $n = 1000$ for both functions with $e \sim N(0, 0.1^2)$ and $e \sim N(0, 2^2)$ for (3.22) and (3.23) respectively. In this simulation, we choose the gird search space of (k_1, k_2) to be $\{1, 2, \ldots, 100\} \otimes \{1, 2, \ldots, 100\}$ for all models. Bandwidths h are selected from the set $\{0.05, 0.055, \ldots, 0.1\}$ for both functions (3.22) and (3.24). The results for second order derivative estimation of function (3.22) and (3.24) are shown in Figure 3.6. For visual purposes the boundary points have been removed. Figure 3.6 shows the second order noisy derivative (blue dots), the true first order derivative (full line) and smoothed first order derivative (red dashed line) for both models (3.22) and (3.24) respectively.



(a) model (3.22) (b) model (3.24)

Figure 3.5: Raw Data for both models

To compare the proposed smoothed second order derivative estimator with the cubic local polynomial estimator, we show both estimators for model (3.22) and model (3.24) in Figure 3.7. It is clear that the proposed second order derivative estimator slightly outperforms the local polynomial ($p = 3$) estimates. For the Monte Carlo study, we construct



(a) model (3.22) (b) model (3.24)

Figure 3.6: Second order derivatives smoothed by $p = 3$ local polynomial regression using a kernel K such that $K(0) = 0$ (red dashed line) on the noisy second order derivative data(blue dots) and true derivative function (full line). (a) Second order derivative of model (3.22), with $k_1 = 42$ and $k_2 = 23$; (b) Second order derivative of model (3.24) with $k_1 = 44$ and $k_2 = 24$. Boundary points are not shown for visual purposes.

data sets of size $n = 700$ for the function

$$m(x) = 8e^{-(1-5x)^3(1-7x)} \quad \text{for} \quad X \sim \mathcal{U}(0,1)$$

100 times according to model (3.2) with $e \sim N(0, 0.1^2)$. As a measure of performance, we define the adjusted mean absolute error as

$$\text{MAEadjusted} = \frac{1}{640} \sum_{i=31}^{670} |\hat{m}_n^{(2)}(X_i) - m^{(2)}(X_i)|$$

to ignore the boundary effects in the simulation result. Bandwidths are selected from interval $\{0.03, 0.035, \ldots, 0.1\}$.

Similar to the first order derivative, we compare the proposed methodology with local polynomial regression (R package `locpol` (Ojeda Cabrera, 2012)) and penalized smoothing splines (R package `stat` (Ramsey and Ripley, 2017)). The order of the local polynomial was taken to be $p = 3$ since p minus the order of the derivative is odd (Fan and Gijbels, 1996). In case of penalized smoothing splines, cubic splines were used. The result is shown in Figure 3.8. The proposed estimator has a slightly better performance compared to local cubic polynomial estimates and penalized cubic splines.

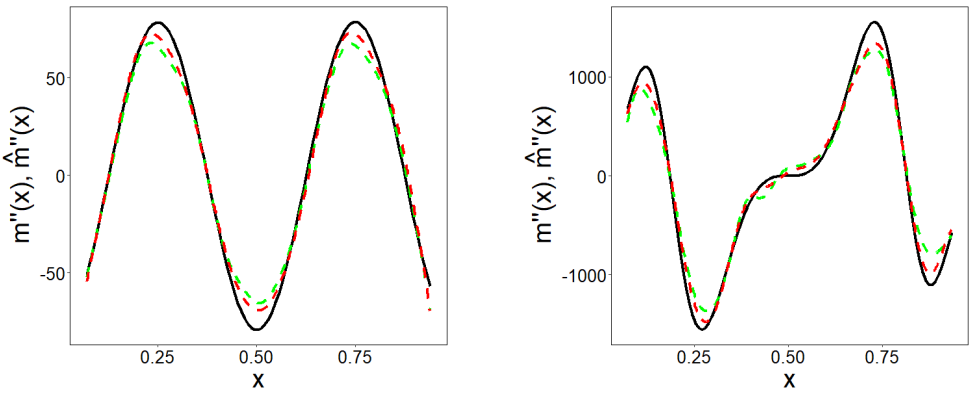


Figure 3.7: Second order derivatives smoothed by $p = 3$ local polynomial regression using a kernel K such that $K(0) = 0$ (red dashed line) on the noisy second order derivative data, the local polynomial estimator with $p = 3$ (green dash line) and true derivative function (full line). (a) Second order derivative of model (3.22) with $k_1 = 42$ and $k_2 = 23$ (b) Second order derivative of model (3.24) with $k_1 = 44$ and $k_2 = 24$. Boundary points are not shown for visual purposes.

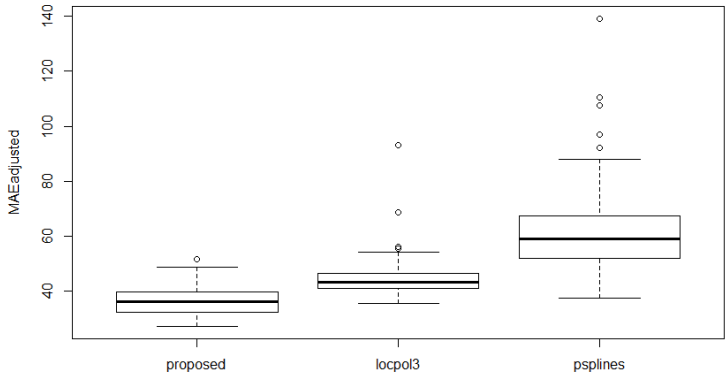


Figure 3.8: Result of the Monte Carlo study for the proposed methodology, local polynomial regression and penalized smoothing splines for second order derivative estimation.

3.5 Conclusions

We proposed a method for derivative estimation in random design and discussed the asymptotic properties of the proposed estimators. The proposed methodology allows to

estimate derivatives nonparametrically without having to estimate the regression function. Asymptotic bias and variance are derived, L_1 and L_2 rates of convergence are established. Our analysis showed that estimating higher order derivatives becomes increasingly more difficult and slower rates of convergence are to be expected. Further, we provide a rule-of-thumb to choose the parameter(s) for the first and second order noisy derivatives. Finally, since the independence assumption of the newly created data set does no longer hold, we use a simple but effective smoothing methodology based on kernels K such that $K(0) = 0$ combined with the flexibility of local polynomial regression. Additionally, we discussed the property of the smoothed noisy derivative estimates.

One drawback of the proposed framework is that the proposed first and second order derivative estimator requires the estimation of the density $f(x)$ and distribution $F(x)$. A first topic of further research interest is to adapt the proposed framework directly for arbitrary distributions without transformation. Second, finding an efficient way to tune h and k simultaneously would greatly benefit the rate of convergence of the proposed methodology. A potential lead could be found in the use of empirical semi-variograms.

3.6 Appendix. Proof of Asymptotic Results for Derivative Estimators

3.6.1 Appendix A. Proof of Lemma 1

Following [p. 14]David and Nagaraja (2003) we have

$$U_{(i+j)} - U_{(i-j)} \sim \text{Beta}(2j, n+1-2j).$$

It immediately follows that

$$\begin{aligned} U_{(i+j)} - U_{(i-j)} &= \mathbf{E}\{U_{(i+j)} - U_{(i-j)}\} + O_p\left[\sqrt{\mathbf{Var}\{U_{(i+j)} - U_{(i-j)}\}}\right] \\ &= \frac{2j}{n+1} + O_p\left(\sqrt{\frac{j}{n^2}}\right) \end{aligned}$$

Similarly, according to the property of uniform order statistics we have

$$U_{(i+j)} - U_{(i)} \sim \text{Beta}(j, n+1-j)$$

and

$$\begin{aligned} U_{(i+j)} - U_{(i)} &= \mathbf{E}\{U_{(i+j)} - U_{(i)}\} + O_p\left[\sqrt{\mathbf{Var}\{U_{(i+j)} - U_{(i)}\}}\right] \\ &= \frac{j}{n+1} + O_p\left(\sqrt{\frac{j}{n^2}}\right). \end{aligned}$$

The proof of the third part of the lemma is analogous to the proof above and is therefore omitted.

3.6.2 Appendix B. Proof of Proposition 1

$$\begin{aligned} \mathbf{Var}\left[\hat{Y}_i^{(1)}|\mathbb{U}\right] &= \mathbf{Var}\left[\sum_{j=1}^{k} w_{i,j} \cdot \left(\frac{Y_{i+j} - Y_{i-j}}{U_{(i+j)} - U_{(i-j)}}\right)|\mathbb{U}\right] \\ &= \left(1 - \sum_{j=2}^{k} w_{i,j}\right)^2 \mathbf{Var}\left[\frac{Y_{i+1} - Y_{i-1}}{U_{(i+1)} - U_{(i-1)}}|\mathbb{U}\right] + \sum_{j=2}^{k} w_{i,j}^2 \mathbf{Var}\left[\frac{Y_{i+j} - Y_{i-j}}{U_{(i+j)} - U_{(i-j)}}|\mathbb{U}\right] \\ &= \left(1 - \sum_{j=2}^{k} w_{i,j}\right)^2 \frac{2\sigma_e^2}{(U_{(i+1)} - U_{(i-1)})^2} + \frac{2\sigma_e^2}{(U_{(i+j)} - U_{(i-j)})^2} \sum_{j=2}^{k} w_{i,j}^2. \end{aligned}$$

Setting the partial derivatives to zero yields

$$w_{i,j} = w_{i,1} \frac{(U_{(i+j)} - U_{(i-j)})^2}{(U_{(i+1)} - U_{(i-1)})^2}.$$

Using the fact that $\sum_{j=1}^{k} w_{i,j} = 1$ results in

$$\sum_{j=1}^{k} w_{i,j} = w_{i,1} \sum_{j=1}^{k} \frac{(U_{(i+j)} - U_{(i-j)})^2}{(U_{(i+1)} - U_{(i-1)})^2} = 1.$$

Consequently, this gives

$$w_{i,j} \frac{(U_{(i+1)} - U_{(i-1)})^2}{(U_{(i+j)} - U_{(i-j)})^2} \sum_{j=1}^{k} \frac{(U_{(i+j)} - U_{(i-j)})^2}{(U_{(i+1)} - U_{(i-1)})^2} = 1$$

proving the proposition.

3.6.3 Appendix C. Proof of Theorem 1

Since r is twice continuously differentiable on $[0,1]$, the following Taylor expansions are valid for $r(U_{(i+j)})$ and $r(U_{(i-j)})$ in a neighborhood of $U_{(i)}$:

$$r(U_{(i+j)}) = r(U_{(i)}) + (U_{(i+j)} - U_{(i)})r'(U_{(i)}) + \frac{(U_{(i+j)} - U_{(i)})^2}{2} r^{(2)}(\zeta_{i,i+j})$$

and

$$r(U_{(i-j)}) = r(U_{(i)}) + (U_{(i-j)} - U_{(i)})r'(U_{(i)}) + \frac{(U_{(i-j)} - U_{(i)})^2}{2} r^{(2)}(\zeta_{i-j,i}),$$

where $\zeta_{i,i+j} \in]U_{(i)}, U_{(i+j)}[$ and $\zeta_{i-j,i} \in]U_{(i-j)}, U_{(i)}[$. Using Lemma 1 and Proposition 1, the absolute conditional bias is bounded above by

$$\begin{aligned}
\left|\text{bias}\left[\hat{Y}_i^{(1)}|\mathbb{U}\right]\right| &= \left|\mathbf{E}\left[\sum_{j=1}^{k} w_{i,j} \cdot \left(\frac{Y_{i+j} - Y_{i-j}}{U_{(i+j)} - U_{(i-j)}}\right)|\mathbb{U}\right] - r'(U_i)\right| \\
&= \frac{1}{2}\left|\sum_{j=1}^{k} w_{i,j} \frac{(U_{(i+j)} - U_{(i)})^2 r^{(2)}(\zeta_{i,i+j}) - (U_{(i-j)} - U_{(i)})^2 r^{(2)}(\zeta_{i-j,i})}{U_{(i+j)} - U_{(i-j)}}\right| \\
&= \frac{1}{2}\Bigg|\frac{\sum_{j=1}^{k}(U_{(i+j)} - U_{(i-j)})(U_{(i+j)} - U_{(i)})^2 r^{(2)}(\zeta_{i,i+j})}{\sum_{l=1}^{k}(U_{(i+l)} - U_{(i-l)})^2} \\
&\quad - \frac{\sum_{j=1}^{k}(U_{(i+j)} - U_{(i-j)})(U_{(i-j)} - U_{(i)})^2 r^{(2)}(\zeta_{i-j,i})}{\sum_{l=1}^{k}(U_{(i+l)} - U_{(i-l)})^2}\Bigg| \\
&\leq \frac{1}{2}\sup_{u\in[0,1]}|r^{(2)}(u)| \frac{\sum_{j=1}^{k}(U_{(i+j)} - U_{(i-j)})\left\{(U_{(i+j)} - U_{(i)})^2 + (U_{(i-j)} - U_{(i)})^2\right\}}{\sum_{l=1}^{k}(U_{(i+l)} - U_{(i-l)})^2} \\
&= \frac{1}{2}\sup_{u\in[0,1]}|r^{(2)}(u)| \frac{\frac{k^2(k+1)^2}{(n+1)^3}\left\{1 + O_p(\frac{1}{\sqrt{k}})\right\}}{\frac{2k(k+1)(2k+1)}{3(n+1)^2}\left\{1 + O_p(\frac{1}{\sqrt{k}})\right\}} \\
&= \sup_{u\in[0,1]}|r^{(2)}(u)| \frac{3k(k+1)}{4(n+1)(2k+1)}\left\{1 + O_p\left(\frac{1}{\sqrt{k}}\right)\right\}.
\end{aligned}$$

Then for $k \to \infty$ as $n \to \infty$

$$\left|\text{bias}\left[\hat{Y}_i^{(1)}|\mathbb{U}\right]\right| \leq \sup_{u\in[0,1]}|r^{(2)}(u)| \frac{3k(k+1)}{4(n+1)(2k+1)}\left\{1 + o_p(1)\right\}$$

Using Proposition 1, the conditional variance yields

$$\begin{aligned}
\mathbf{Var}\big[\hat{Y}_i^{(1)}|\mathbb{U}\big] &= \mathbf{Var}\left[\sum_{j=1}^{k} w_{i,j}\cdot\left(\frac{Y_{i+j}-Y_{i-j}}{U_{(i+j)}-U_{(i-j)}}\right)|\mathbb{U}\right] \\
&= 2\sigma_e^2\frac{\sum_{j=1}^{k}(U_{(i+j)}-U_{(i-j)})^2}{\left(\sum_{l=1}^{k}(U_{(i+l)}-U_{(i-l)})^2\right)^2} \\
&= 2\sigma_e^2\frac{1}{\sum_{l=1}^{k}\left(U_{(i+l)}-U_{(i-l)}\right)^2} \\
&= 2\sigma_e^2\frac{1}{\frac{2k(k+1)(2k+1)}{3(n+1)^2}\{1+o_p(1)\}} \\
&= \frac{3\sigma_e^2(n+1)^2}{k(k+1)(2k+1)}\{1+o_p(1)\},
\end{aligned}$$

provided that $k\to\infty$ as $n\to\infty$. Both results hold uniformly for $k+1\le i\le n-k$.

3.6.4 Appendix D. Proof of Corollary 1

Under the conditions $k\to\infty$ as $n\to\infty$ such that $n^{-1}k\to 0$ and $n^2k^{-3}\to 0$, Theorem 1 states that the upperbound of conditional bias and conditional variance go to zero. Consequently, we have that

$$\lim_{n\to\infty}\mathrm{MSE}\big[\hat{Y}_i^{(1)}|\mathbb{U}\big]=0.$$

According to Chebyshev's inequality the proof is complete.

3.6.5 Appendix E. Proof of Corollary 2

From the bias-variance decomposition of the mean squared error (MSE), it follows that

$$\mathrm{MSE}\big[\hat{Y}_i^{(1)}|\mathbb{U}\big]\le\mathcal{B}^2\frac{9k^2(k+1)^2}{16(n+1)^2(2k+1)^2}+\frac{3\sigma_e^2(n+1)^2}{k(k+1)(2k+1)}+o_p(n^{-2}k^2+n^2k^{-3}),$$

with $\mathcal{B}=\sup_{u\in[0,1]}|r^{(2)}(u)|$. Since $U\sim\mathcal{U}(0,1)$, the conditional mean integrated squared error (MISE) which measures the average global error is:

$$\begin{aligned}
\mathrm{MISE}\big[\hat{Y}^{(1)}|\mathbb{U}\big] &= \mathbf{E}\int_0^1\big[\hat{Y}^{(1)}(U)-r^{(1)}(U)|\mathbb{U}\big]^2dU \\
&= \int_0^1\mathbf{E}\big[\hat{Y}^{(1)}(U)-r^{(1)}(U)|\mathbb{U}\big]^2dU \\
&\le B^2\frac{9k^2(k+1)^2}{16(n+1)^2(2k+1)^2}+\frac{3\sigma_e^2(n+1)^2}{k(k+1)(2k+1)}+o_p(n^{-2}k^2+n^2k^{-3})
\end{aligned}$$

with $\hat{Y}^{(1)}(U)$ represents the first order derivative estimator at design point U. Denote the asymptotic conditional MISE (AMISE) by

$$\text{AMISE}\big[\hat{Y}^{(1)}|\mathbb{U}\big] \leq \mathcal{B}^2 \frac{9k^2(k+1)^2}{16(n+1)^2(2k+1)^2} + \frac{3\sigma_e^2(n+1)^2}{k(k+1)(2k+1)}.$$

3.6.6 Appendix F. Proof of Exact Bias

Assume the $q+1$ derivatives of r exist on $[0,1]$, according to lemma 1, the following Taylor expansions are valid for $r(U_{(i+j)})$ and $r(U_{(i-j)})$ in a neighborhood of $U_{(i)}$

$$\begin{aligned} r(U_{(i+j)}) &= r(U_{(i)}) + \sum_{l=1}^{q} \frac{1}{l!}(U_{(i+j)} - U_{(i)})^l r^{(l)}(U_{(i)}) + O_p(U_{(i+j)} - U_{(i)})^{q+1} \\ &= r(U_{(i)}) + \sum_{l=1}^{q} \frac{1}{l!}(U_{(i+j)} - U_{(i)})^l r^{(l)}(U_{(i)}) + O_p\{(j/n)^{q+1}\} \end{aligned}$$

and

$$\begin{aligned} r(U_{(i-j)}) &= r(U_{(i)}) + \sum_{l=1}^{q} \frac{1}{l!}(U_{(i-j)} - U_{(i)})^l r^{(l)}(U_{(i)}) + O_p(U_{(i-j)} - U_{(i)})^{q+1} \\ &= r(U_{(i)}) + \sum_{l=1}^{q} \frac{1}{l!}(U_{(i-j)} - U_{(i)})^l r^{(l)}(U_{(i)}) + O_p\{(j/n)^{q+1}\}. \end{aligned}$$

Taking expectations and for $\sum_{j=1}^{k} w_{i,j} = 1$

$$\begin{aligned} \mathbf{E}\big[\hat{Y}_i^{(1)}|\mathbb{U}\big] &= \sum_{j=1}^{k} w_{i,j} \frac{r(U_{(i+j)}) - r(U_{(i-j)})}{U_{(i+j)} - U_{(i-j)}} \\ &= \frac{1}{\sum_{p=1}^{k}(U_{(i+p)} - U_{(i-p)})^2} \Bigg(\sum_{j=1}^{k} (U_{(i+j)} - U_{(i-j)}) \Bigg[\sum_{l=1}^{q} \frac{r^{(l)}(U_{(i)})}{l!} \big\{ (U_{(i+j)} - U_{(i)})^l \\ &- (U_{(i-j)} - U_{(i)})^l \big\} + O_p\{(j/n)^{q+1}\} \Bigg] \Bigg) \end{aligned}$$

For $q = 1$, the second order derivative of r exists on $[0,1]$

$$\begin{aligned} \text{bias}\big[\hat{Y}_i^{(1)}|\mathbb{U}\big] &= \frac{r^{(1)}(U_{(i)}) \sum_{j=1}^{k}(U_{(i+j)} - U_{(i-j)})^2 + O_p\left(k^4/n^3\right)}{\sum_{p=1}^{k}(U_{(i+p)} - U_{(i-p)})^2} - r^{(1)}(U_{(i)}) \\ &= O_p\left(\frac{k}{n}\right) \end{aligned}$$

For $q = 2$, the third order derivative of r exists on $[0, 1]$

$$\begin{aligned}
\text{bias}\big[\hat{Y}_i^{(1)}|\mathbb{U}\big] &= \frac{r^{(2)}(U_{(i)})}{2\sum_{p=1}^{k}(U_{(i+p)} - U_{(i-p)})^2}\Big[\sum_{j=1}^{k}(U_{(i+j)} - U_{(i-j)})\big\{(U_{(i+j)} - U_{(i)})^2 \\
&- (U_{(i-j)} - U_{(i)})^2\big\} + O_p\left(k^5/n^4\right)\Big] \\
&= \frac{O_p(k^{\frac{7}{2}}/n^3) + O_p\left(k^5/n^4\right)}{O_p(k^3/n^2)} \\
&= O_p\left(\max\left\{\frac{k^{\frac{1}{2}}}{n}, \frac{k^2}{n^2}\right\}\right)
\end{aligned}$$

The bias can be split into two terms, $\text{bias}_{\text{even}} = O_p\left(\frac{k^{\frac{1}{2}}}{n}\right)$ and $\text{bias}_{\text{odd}} = O_p\left(\frac{k^2}{n^2}\right)$. $\text{bias}_{\text{even}}$ indicates the bias from the even order terms in the Taylor expansion of $r(U_{(i\pm j)})$ and bias_{odd} for the odd order terms respectively.

For $q > 2$, we have

$$\begin{aligned}
\mathbf{E}\big[\hat{Y}_i^{(1)}|\mathbb{U}\big] &= r^{(1)}(U_{(i)}) + \frac{1}{\sum_{p=1}^{k}(U_{(i+p)} - U_{(i-p)})^2}\Big(\sum_{j=1}^{k}(U_{(i+j)} - U_{(i-j)}) \\
&\Big[\sum_{l=2}^{q}\frac{r^{(l)}(U_{(i)})}{l!}\big\{(U_{(i+j)} - U_{(i)})^l - (U_{(i-j)} - U_{(i)})^l\big\} + O_p\{(j/n)^{q+1}\}\Big]\Big).
\end{aligned}$$

Splitting the second term in $\text{bias}_{\text{even}}$ and $\text{bias}_{\text{even}}$ yields

$$\begin{aligned}
\text{bias}_{\text{odd}}\big[\hat{Y}_i^{(1)}|\mathbb{U}\big] &\stackrel{\text{def}}{=} \frac{1}{\sum_{p=1}^{k}(U_{(i+p)} - U_{(i-p)})^2}\Big(\sum_{j=1}^{k}(U_{(i+j)} - U_{(i-j)}) \\
&\Big[\sum_{l=3,5,\ldots}^{2\lceil q/2\rceil - 1}\frac{r^{(l)}(U_{(i)})}{l!}\Big((U_{(i+j)} - U_{(i)})^l - (U_{(i-j)} - U_{(i)})^l\Big)\Big]\Big) \\
&= \frac{\sum_{j=1}^{k}(U_{(i+j)} - U_{(i-j)})\left[\frac{r^{(3)}(U_i)}{6}\Big((U_{(i+j)} - U_{(i)})^3 - (U_{(i-j)} - U_{(i)})^3\Big)\right]}{\sum_{p=1}^{k}(U_{(i+p)} - U_{(i-p)})^2} \\
&\quad \big\{1 + o_p(1)\big\} \\
&= \left(\frac{r^{(3)}(U_{(i)})(3k^2 + 3k - 1)}{30(n+1)^2} + O_p\left(\frac{k^{\frac{3}{2}}}{n^2}\right)\right)\big\{1 + o_p(1)\big\} \\
&= O_p\left\{\frac{k^2}{n^2}\right\}
\end{aligned}$$

$$
\begin{aligned}
\mathrm{bias}_{\mathrm{even}}[\hat{Y}_i^{(1)}|\mathbb{U}] &\overset{\mathrm{def}}{=} \frac{1}{\sum_{p=1}^{k}(U_{(i+p)}-U_{(i-p)})^2}\Bigg(\sum_{j=1}^{k}(U_{(i+j)}-U_{(i-j)}) \\
&\quad \left[\sum_{l=2,4,\ldots}^{2\lfloor q/2\rfloor}\frac{r^{(l)}(U_{(i)})}{l!}\left((U_{(i+j)}-U_{(i)})^l-(U_{(i-j)}-U_{(i)})^l\right)\right]\Bigg) \\
&= \frac{\sum_{j=1}^{k}(U_{(i+j)}-U_{(i-j)})\left[\frac{r^{(2)}(U_{(i)})}{2}\left((U_{(i+j)}-U_{(i)})^2-(U_{(i-j)}-U_{(i)})^2\right)\right]}{\sum_{p=1}^{k}(U_{(i+p)}-U_{(i-p)})^2} \\
&\quad \{1+o_p(1)\} \\
&= O_p\left\{\frac{k^{\frac{1}{2}}}{n}\right\}
\end{aligned}
$$

3.6.7 Appendix G. Bias and Variance at the Left Boundary

Assume that r is three times continuously differentiable on $[0,1]$. At the left boundary $i < k+1$, we have

$$
\begin{aligned}
&\mathrm{bias}[\hat{Y}_i^{(1)}|\mathbb{U}] = \sum_{j=1}^{k(i)} w_{i,j}\cdot\left(\frac{\frac{1}{2}\left[(U_{(i+j)}-U_{(i)})^2-\frac{1}{2}(U_{(i-j)}-U_{(i)})^2\right]r^{(2)}(U_{(i)})}{U_{(i+j)}-U_{(i-j)}}\right) \\
&+\sum_{j=1}^{k(i)} w_{i,j}\cdot\left(\frac{O_p(j^3/n^3)}{U_{(i+j)}-U_{(i-j)}}\right)+\sum_{j=k(i)+1}^{k} w_{i,j}\cdot\left(\frac{1}{2}(U_{(i+j)}-U_{(i)})r^{(2)}(U_{(i)})\right)\{1+o_p(1)\} \\
&= O_p\left\{\max\left(\frac{k(i)^{7/2}}{k^3 n},\frac{k(i)^5}{k^3 n^2},\frac{k-k(i)}{n}\right)\right\} \\
&\mathbf{Var}[\hat{Y}_i^{(1)}|\mathbb{U}] = \mathbf{Var}\left[\sum_{j=1}^{k(i)} w_{i,j}\left(\frac{Y_{i+j}-Y_{i-j}}{U_{(i+j)}-U_{(i-j)}}\right)|\mathbb{U}\right]+\mathbf{Var}\left[\sum_{j=k(i)+1}^{k} w_{i,j}\left(\frac{Y_{i+j}-Y_i}{U_{(i+j)}-U_{(i)}}\right)|\mathbb{U}\right] \\
&= 2\sigma_e^2\sum_{j=1}^{k(i)}\left(\frac{w_{i,j}}{U_{(i+j)}-U_{(i-j)}}\right)^2+\sigma_e^2\sum_{j=k(i)+1}^{k}\left(\frac{w_{i,j}}{U_{(i+j)}-U_{(i)}}\right)^2 \\
&+ \sigma_e^2\left[\sum_{j=k(i)+1}^{k}\left(\frac{w_{i,j}}{U_{(i+j)}-U_{(i)}}\right)\right]^2 \\
&= O_p\left\{\max\left(\frac{n^2}{k^3},\frac{n^2(k-k(i))^2}{k^4}\right)\right\}.
\end{aligned}
$$

3.6.8 Appendix H. Proof of Theorem 2

Part I (conditional bias)

$$\begin{aligned}
\text{bias}\big[\hat{r}^{(1)}(u_0)|\widetilde{\mathbb{U}}\big] &= \mathbf{E}\big[\hat{r}^{(1)}(u_0)|\widetilde{\mathbb{U}}\big] - r^{(1)}(u_0) \\
&= \boldsymbol{\epsilon}_1^T\, \mathbf{S}_n^{-1}\, \mathbf{U}_{\mathrm{u}}{}^T\, \mathbf{W}_{\mathrm{u}}\, \mathbf{E}[\hat{\mathbf{Y}}^{(1)}|\widetilde{\mathbb{U}}] - r^{(1)}(u_0) \\
&= \boldsymbol{\epsilon}_1^T\, \mathbf{S}_n^{-1}\, \mathbf{U}_{\mathrm{u}}{}^T\, \mathbf{W}_{\mathrm{u}} \left(\begin{bmatrix} r^{(1)}(U_{(k+1)}) \\ \vdots \\ r^{(1)}(U_{(n-k)}) \end{bmatrix} + \begin{bmatrix} \text{bias}[\hat{Y}^{(1)}_{k+1}|\mathbb{U}] \\ \vdots \\ \text{bias}[\hat{Y}^{(1)}_{n-k}|\mathbb{U}] \end{bmatrix} \right) - r^{(1)}(u_0)
\end{aligned}$$

For p odd, (see Theorem 3.1 in Fan and Gijbels (1996)), the first term is

$$\begin{aligned}
\boldsymbol{\epsilon}_1^T\, \mathbf{S}_n^{-1}\, \mathbf{U}_{\mathrm{u}}{}^T\, \mathbf{W}_{\mathrm{u}} \begin{bmatrix} r^{(1)}(U_{(k+1)}) \\ \vdots \\ r^{(1)}(U_{(n-k)}) \end{bmatrix} - r^{(1)}(u_0) &= \boldsymbol{\epsilon}_1^T\, \mathbf{S}^{-1}\, c_p \beta_{p+1} h^{p+1} + o_p(h^{p+1}) \\
&= \boldsymbol{\epsilon}_1^T\, \mathbf{S}^{-1}\, \frac{c_p}{(p+1)!} r^{(p+2)}(u_0) h^{p+1} + o_p(h^{p+1})
\end{aligned} \tag{3.25}$$

where $c_p = (\mu_{p+1}, \ldots, \mu_{2p+1})^T$ with $\mu_j = \int u^j K(u) du$, and $\mathbf{S} = (\mu_{i+j})_{0 \le i,j \le p}$.

Based on Theorem 1, for $k \to \infty$ as $n \to \infty$ the second term is

$$\begin{aligned}
&\boldsymbol{\epsilon}_1^T\, \mathbf{S}_n^{-1}\, \mathbf{U}_{\mathrm{u}}{}^T\, \mathbf{W}_{\mathrm{u}} \begin{bmatrix} \text{bias}[\hat{Y}^{(1)}_{k+1}|\mathbb{U}] \\ \vdots \\ \text{bias}[\hat{Y}^{(1)}_{n-k}|\mathbb{U}] \end{bmatrix} \\
&\le \boldsymbol{\epsilon}_1^T\, \mathbf{S}_n^{-1}\, \mathbf{U}_{\mathrm{u}}{}^T\, \mathbf{W}_{\mathrm{u}} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \sup_{u \in [0,1]} |r^{(2)}(u)| \frac{3k(k+1)}{4(n+1)(2k+1)} \{1 + o_p(1)\}.
\end{aligned} \tag{3.26}$$

Ignore orders statistics among $U_{(k+1)}, \ldots, U_{(n-k)}$, they can be treated i.i.d samples. Let $S_{n-2k,l} = \sum_{m=k+1}^{n-k} (U_{(m)} - u_0)^l K_h(U_{(m)} - u_0)$, then for $l = 0, 1, \ldots, p$

$$S_{n-2k,l} = \mathbf{E}[S_{n-2k,l}] + O_p\left\{ \sqrt{\mathbf{Var}[S_{n-2k,l}]} \right\}$$

For $h \to 0$ and $nh \to \infty$ as $n \to \infty$ we have

$$\begin{aligned}
\mathbf{E}[S_{n-2k,l}] &= (n-2k)\,\mathbf{E}[(U-u_0)^l K_h(U-u_0)] \\
&= \frac{(n-2k)}{h}\int K\left(\frac{u-u_0}{h}\right)(u-u_0)^l f(u)du \\
&= (n-2k)h^l \int K(x)x^l f(u_0+xh)dx \\
&= (n-2k)h^l f(u_0)\left[\int x^l K(x)dx + o_p(1)\right] \\
&= (n-2k)h^l f(u_0)\mu_l\{1+o_p(1)\}
\end{aligned}$$

and similarly

$$\begin{aligned}
O_p\left\{\sqrt{\mathbf{Var}[S_{n-2k,l}]}\right\} &= O_p\left(\sqrt{(n-2k)\,\mathbf{E}[(U-u_0)^{2l}K_h^2(U-u_0)]}\right) \\
&= O_p\left(\sqrt{(n-2k)\int (u-u_0)^{2l}K_h^2(u-u_0)f(u)du}\right) \\
&= O_p\left(\sqrt{(n-2k)h^{2l-1}f(u_0)\int x^{2l}K^2(x)dx}\right) \\
&= O_p\left(\sqrt{(n-2k)h^{2l-1}}\right)
\end{aligned}$$

Thus for $h \to 0$ and $nh \to \infty$ as $n \to \infty$, $k \to \infty$ as $n \to \infty$ such that $n^{-1}k \to 0$:

$$\begin{aligned}
S_{n-2k,l} &= (n-2k)h^l f(u_0)\mu_l\left\{1+o_p(1)+O_p\left(\sqrt{1/(h(n-2k))}\right)\right\} \\
&= (n-2k)h^l f(u_0)\mu_l\left\{1+o_p(1)\right\} \qquad (3.27)
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{S}_n &= \mathbf{U}_{\mathrm{u}}{}^T\,\mathbf{W}_{\mathrm{u}}\,\mathbf{U}_{\mathrm{u}} = \begin{bmatrix} S_{n-2k,0} & S_{n-2k,1} & \cdots & S_{n-2k,p} \\ S_{n-2k,1} & S_{n-2k,1} & \cdots & S_{n-2k,p+1} \\ \vdots & \vdots & \ddots & \vdots \\ S_{n-2k,p} & S_{n-2k,p+1} & \cdots & S_{n-2k,2p} \end{bmatrix} \\
&= (n-2k)f(u_0)H\,\mathbf{S}\,H\{1+o_p(1)\} \qquad (3.28)
\end{aligned}$$

where $H = diag\{1, h, \ldots, h^p\}$. Next,

$$\begin{aligned}
\mathbf{U_u}^T \mathbf{W_u} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} &= \begin{bmatrix} \sum_{m=k+1}^{n-k} K_h(U_{(m)} - u_0) \\ \sum_{m=k+1}^{n-k} (U_{(m)} - u) K_h(U_{(m)} - u_0) \\ \vdots \\ \sum_{m=k+1}^{n-k} (U_{(m)} - u)^p K_h(U_{(m)} - u_0) \end{bmatrix} = \begin{bmatrix} S_{n-2k,0} \\ S_{n-2k,1} \\ \vdots \\ S_{n-2k,p} \end{bmatrix} \\
&= (n-2k) f(u_0) H \tilde{c}_p \{1 + o_p(1)\} \qquad (3.29)
\end{aligned}$$

where $\tilde{c}_p = (\mu_0, \mu_1, \ldots, \mu_p)^T$. Plugging (3.28),(3.29) into (3.26) gives

$$\begin{aligned}
\boldsymbol{\epsilon}_1^T \mathbf{S}_n^{-1} \mathbf{U_u}^T \mathbf{W_u} \begin{bmatrix} \text{bias}[\hat{Y}_{k+1}^{(1)} | \mathbb{U}] \\ \vdots \\ \text{bias}[\hat{Y}_{n-k}^{(1)} | \mathbb{U}] \end{bmatrix} & \\
\leq \sup_{u \in [0,1]} |r^{(2)}(u)| \frac{3k(k+1)}{4(n+1)(2k+1)} \boldsymbol{\epsilon}_1^T H^{-1} \mathbf{S}^{-1} \tilde{c}_p \{1 + o_p(1)\} & \\
= \sup_{u \in [0,1]} |r^{(2)}(u)| \frac{3k(k+1)}{4(n+1)(2k+1)} \boldsymbol{\epsilon}_1^T \mathbf{S}^{-1} \tilde{c}_p \{1 + o_p(1)\} & \qquad (3.30)
\end{aligned}$$

Based on (3.25) and (3.30), we have

$$\begin{aligned}
&\text{bias}\left[\hat{r}^{(1)}(u_0) | \widetilde{\mathbb{U}}\right] \\
&= \boldsymbol{\epsilon}_1^T \mathbf{S}_n^{-1} \mathbf{U_u}^T \mathbf{W_u} \begin{bmatrix} r^{(1)}(U_{(k+1)}) \\ \vdots \\ r^{(1)}(U_{(n-k)}) \end{bmatrix} - r^{(1)}(u_0) + \boldsymbol{\epsilon}_1^T \mathbf{S}_n^{-1} \mathbf{U_u}^T \mathbf{W_u} \begin{bmatrix} \text{bias}[\hat{Y}_{k+1}^{(1)} | \mathbb{U}] \\ \vdots \\ \text{bias}[\hat{Y}_{n-k}^{(1)} | \mathbb{U}] \end{bmatrix} \\
&\leq \boldsymbol{\epsilon}_1^T \mathbf{S}^{-1} \left[\frac{c_p}{(p+1)!} r^{(p+2)}(u_0) h^{p+1} + \sup_{u \in [0,1]} |r^{(2)}(u)| \frac{3k(k+1)}{4(n+1)(2k+1)} \tilde{c}_p \right] \{1 + o_p(1)\}.
\end{aligned} \qquad (3.31)$$

Part II (conditional variance)

Provided that $k \to \infty$ as $n \to \infty$, consider the conditional variance in Theorem 1

$$\sigma_{\tilde{e}^2} = \frac{3\sigma_e^2 (n+1)^2}{k(k+1)(2k+1)} \{1 + o_p(1)\}$$

and by Theorem 1 in De Brabanter et al. (2018)

$$\begin{aligned}\mathbf{Var}\big[\hat{r}^{(1)}(u_0)|\widetilde{\mathbb{U}}\big] &= \boldsymbol{\epsilon}_1^T\,\mathbf{S}_n^{-1}(\mathbf{U}_{\mathrm{u}}{}^T\,\mathbf{W}_{\mathrm{u}}\,\mathbf{Var}[\hat{\mathbf{Y}}^{(1)}|\widetilde{\mathbb{U}}]\,\mathbf{W}_{\mathrm{u}}\,\mathbf{U}_{\mathrm{u}})\,\mathbf{S}_n^{-1}\,\boldsymbol{\epsilon}_1 \\ &= \frac{3\sigma_e^2(n+1)^2}{k(k+1)(2k+1)}\frac{1+f(u_0)\rho_c}{h(n-2k)f(u_0)}\boldsymbol{\epsilon}_1^T\,\mathbf{S}^{-1}\,\mathbf{S}^*\,\mathbf{S}^{-1}\,\boldsymbol{\epsilon}_1\{1+o_p(1)\} \qquad (3.32)\end{aligned}$$

with $\lim_{n\to\infty} n\int \rho_n(x)dx = \rho_c$ and $\mathbf{S}^* = (\nu_{i+j})_{0\le i,j\le p}$ with $\nu_j = \int u^j K^2(u)du$.

For p odd, (see Theorem 3.1 in Fan and Gijbels (1996)):

$$\begin{aligned}\int K_0^*(t)dt &= \boldsymbol{\epsilon}_1^T S^{-1}\left(\int K(t)dt, \int tK(t)dt, \ldots, \int t^p K(t)dt\right)^T \\ &= \boldsymbol{\epsilon}_1^T S^{-1}\tilde{c}_p \qquad (3.33)\end{aligned}$$

Similarly, we obtain

$$\int t^{p+1}K_0^*(t)dt = \boldsymbol{\epsilon}_1^T S^{-1}c_p, \quad \int K_0^{*2}(t)dt = \boldsymbol{\epsilon}_1^T S^{-1}S^*S^{-1}\boldsymbol{\epsilon}_1 \qquad (3.34)$$

plugging (3.33) and (3.34) into (3.31) and (3.32) provides the second part of Theorem 2.

3.6.9 Appendix I. Proof of Corollary 3

For $h \to 0$, $nh \to \infty$ and $k \to \infty$ as $n \to \infty$ such that $n^{-1}k \to 0$ and $nk^{-3}h^{-1} \to 0$, then theorem 2 states that the upperbound of the conditional bias and conditional variance go to zero. Consequently, we have that

$$\lim_{n\to\infty} \mathrm{MSE}\big[\hat{r}^{(1)}(u_0)|\widetilde{\mathbb{U}}\big] = 0.$$

According to Chebyshev's inequality the proof is complete.

3.6.10 Appendix J. Proof of Theorem 3

The proof for the asymptotic properties of the second order derivatives is similar to that of the first order derivatives. Since r is three times continuously differentiable on the compact interval $[0,1]$, the following Taylor expansions are valid for $r(U_{(i+j+k_1)})$ and

$r(U_{(i-j-k_1)})$ in a neighborhood of $U_{(i+j)}$ and $U_{(i-j)}$ respectively

$$\begin{aligned} r(U_{(i+j+k_1)}) &= r(U_{(i+j)}) + \sum_{q=1}^{2} \frac{1}{q!}(U_{(i+j+k_1)} - U_{(i+j)})^q r^{(q)}(U_{(i+j)}) \\ &+ \frac{(U_{(i+j+k_1)} - U_{(i+j)})^3}{6} r^{(3)}(\zeta_{i+j,i+j+k_1}) \end{aligned}$$

and

$$\begin{aligned} r(U_{(i-j-k_1)}) &= r(U_{(i-j)}) + \sum_{q=1}^{2} \frac{1}{q!}(U_{(i-j-k_1)} - U_{(i-j)})^q r^{(q)}(U_{(i-j)}) \\ &+ \frac{(U_{(i-j-k_1)} - U_{(i-j)})^3}{6} r^{(3)}(\zeta_{i-j-k_1,i-j}), \end{aligned}$$

where $\zeta_{i+j,i+j+k_1} \in]U_{(i+j)}, U_{(i+j+k_1)}[$ and $\zeta_{i-j-k_1,i-j} \in]U_{(i-j-k_1)}, U_{(i-j)}[$.

Similarly, the following Taylor expansions are also valid for $r^{(1)}(U_{(i+j)})$ and $r^{(1)}(U_{(i-j)})$ in a neighborhood of $U_{(i)}$:

$$r^{(1)}(U_{(i+j)}) = r^{(1)}(U_{(i)}) + (U_{(i+j)} - U_{(i)}) r^{(2)}(U_{(i)}) + \frac{(U_{(i+j)} - U_{(i)})^2}{2} r^{(3)}(\zeta_{i,i+j})$$

and

$$r^{(1)}(U_{(i-j)}) = r^{(1)}(U_{(i)}) + (U_{(i-j)} - U_{(i)}) r^{(2)}(U_{(i)}) + \frac{(U_{(i-j)} - U_{(i)})^2}{2} r^{(3)}(\zeta_{i-j,i}),$$

where $\zeta_{i,i+j} \in]U_{(i)}, U_{(i+j)}[$ and $\zeta_{i-j,i} \in]U_{(i-j)}, U_{(i)}[$.

$$\begin{aligned} r^{(2)}(U_{(i+j)}) &= r^{(2)}(U_{(i)}) + (U_{(i+j)} - U_{(i)}) r^{(3)}(\zeta'_{i,i+j}) \\ r^{(2)}(U_{(i-j)}) &= r^{(2)}(U_{(i)}) + (U_{(i-j)} - U_{(i)}) r^{(3)}(\zeta'_{i-j,i}), \end{aligned}$$

where $\zeta'_{i,i+j} \in]U_{(i)}, U_{(i+j)}[$ and $\zeta'_{i-j,i} \in]U_{(i-j)}, U_{(i)}[$. Since $\sum_{j=1}^{k_2} w_{i,j,2} = 1$, the absolute conditional bias is

$$
\begin{aligned}
&\left|\text{bias}\left[\hat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\right]\right| = \left|\mathbf{E}\left[\hat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\right] - r^{(2)}(U_{(i)})\right| \\
&= \left|2\sum_{j=1}^{k_2} w_{i,j,2} \frac{\left(\frac{r(U_{(i+j+k_1)})-r(U_{(i+j)})}{U_{(i+j+k_1)}-U_{(i+j)}} - \frac{r(U_{(i-j-k_1)})-r(U_{(i-j)})}{U_{(i-j-k_1)}-U_{(i-j)}}\right)}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}} - r^{(2)}(U_{(i)})\right| \\
&= \left|2\sum_{j=1}^{k_2} w_{i,j,2}\left\{\frac{\left(r^{(1)}(U_{(i+j)}) - r^{(1)}(U_{(i-j)}) + \frac{1}{2}r^{(2)}(U_{(i+j)})(U_{(i+j+k_1)}-U_{(i+j)})\right)}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}}\right. \\
&+ \frac{-\frac{1}{2}r^{(2)}(U_{(i-j)})(U_{(i-j-k_1)}-U_{(i-j)}) + \frac{1}{6}r^{(3)}(\zeta_{i+j,i+j+k_1})(U_{(i+j+k_1)}-U_{(i+j)})^2}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}} \\
&\left.+ \frac{-\frac{1}{6}r^{(3)}(\zeta_{i-j-k_1,i-j})(U_{(i-j-k_1)}-U_{(i-j)})^2}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}}\right\} - r^{(2)}(U_{(i)})\Bigg| \\
&= \left|2\sum_{j=1}^{k_2} w_{i,j,2}\left\{\frac{\frac{1}{2}r^{(3)}(\zeta_{i,i+j})(U_{(i+j)}-U_{(i)})^2 - \frac{1}{2}r^{(3)}(\zeta_{i-j,i})(U_{(i-j)}-U_{(i)})^2}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}}\right. \right. \\
&+ \frac{\frac{1}{2}r^{(3)}(\zeta'_{i,i+j})(U_{(i+j)}-U_{(i)})(U_{(i+j+k_1)}-U_{(i+j)})}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}} \\
&- \frac{\frac{1}{2}r^{(3)}(\zeta'_{i-j,i})(U_{(i-j)}-U_{(i)})(U_{(i-j-k_1)}-U_{(i-j)})}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}} \\
&\left.\left.+ \frac{\frac{1}{6}r^{(3)}(\zeta_{i+j,i+j+k_1})(U_{(i+j+k_1)}-U_{(i+j)})^2 - \frac{1}{6}r^{(3)}(\zeta_{i-j-k_1,i-j})(U_{(i-j-k_1)}-U_{(i-j)})^2}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}}\right\}\right| \\
&\leq \sup_{u\in[0,1]}|r^{(3)}(u)|\left(\sum_{j=1}^{k_2} w_{i,j,2}\frac{(U_{(i+j)}-U_{(i)})^2 + (U_{(i-j)}-U_{(i)})^2}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}}\right. \\
&+ \sum_{j=1}^{k_2} w_{i,j,2}\frac{(U_{(i+j)}-U_{(i)})(U_{(i+j+k_1)}-U_{(i+j)}) + (U_{(i-j)}-U_{(i)})(U_{(i-j-k_1)}-U_{(i-j)})}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}} \\
&\left.+ \sum_{j=1}^{k_2} w_{i,j,2}\frac{\frac{1}{3}(U_{(i+j+k_1)}-U_{(i+j)})^2 + \frac{1}{3}(U_{(i-j-k_1)}-U_{(i-j)})^2}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}}\right)
\end{aligned}
$$

Using Lemma 1, the weights in Equation (3.16), $k_1 \to \infty$ and $k_2 \to \infty$ as $n \to \infty$ gives

$$
\left|\text{bias}\left[\hat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\right]\right| \leq \frac{\sup_{u\in[0,1]}|r^{(3)}(u)|}{n+1}\frac{2\sum_{j=1}^{k_2} j^3 + 3k_1\sum_{j=1}^{k_2} j^2 + \frac{5}{3}k_1^2\sum_{j=1}^{k_2} j + \frac{1}{3}k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1\sum_{j=1}^{k_2} j}\{1+o_p(1)\}
$$

Using the weights in Equation (3.16) and Lemma 1, the conditional variance is

$$\begin{aligned}
&\mathbf{Var}\big[\hat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\big] = \mathbf{Cov}\big[2\sum_{j=1}^{k_2} w_{i,j,2}\frac{\left(\frac{Y_{i+j+k_1}-Y_{i+j}}{U_{(i+j+k_1)}-U_{(i+j)}} - \frac{Y_{i-j-k_1}-Y_{i-j}}{U_{(i-j-k_1)}-U_{(i-j)}}\right)}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}},\\
&2\sum_{l=1}^{k_2} w_{i,l,2}\frac{\left(\frac{Y_{i+l+k_1}-Y_{i+l}}{U_{(i+l+k_1)}-U_{(i+l)}} - \frac{Y_{i-l-k_1}-Y_{i-l}}{U_{(i-l-k_1)}-U_{(i-l)}}\right)}{U_{(i+l+k_1)}+U_{(i+l)}-U_{(i-l-k_1)}-U_{(i-l)}}|\widetilde{\mathbb{U}}\big]\\
&= 4\sum_{j=1}^{k_2}\sum_{l=1}^{k_2}\frac{w_{i,j,2}w_{i,l,2}}{(U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)})(U_{(i+l+k_1)}+U_{(i+l)}-U_{(i-l-k_1)}-U_{(i-l)})}\\
&\Bigg\{\frac{\mathbf{Cov}\big[Y_{i+j+k_1}-Y_{i+j},Y_{i+l+k_1}-Y_{i+l}\big]}{(U_{(i+j+k_1)}-U_{(i+j)})(U_{(i+l+k_1)}-U_{(i+l)})} - \frac{\mathbf{Cov}\big[Y_{i+j+k_1}-Y_{i+j},Y_{i-l-k_1}-Y_{i-l}\big]}{(U_{(i+j+k_1)}-U_{(i+j)})(U_{(i-l-k_1)}-U_{(i-l)})}\\
&- \frac{\mathbf{Cov}\big[Y_{i-j-k_1}-Y_{i-j},Y_{i+l+k_1}-Y_{i+l}\big]}{(U_{(i-j-k_1)}-U_{(i-j)})(U_{(i+l+k_1)}-U_{(i+l)})} + \frac{\mathbf{Cov}\big[Y_{i-j-k_1}-Y_{i-j},Y_{i-l-k_1}-Y_{i-l}\big]}{(U_{(i-j-k_1)}-U_{(i-j)})(U_{(i-l-k_1)}-U_{(i-l)})}\Bigg\}\\
&= 4\sigma^2\sum_{j=1}^{k_2}\frac{w_{i,j,2}^2}{(U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)})^2}\left(\frac{2}{(U_{(i+j+k_1)}-U_{(i+j)})^2}+\frac{2}{(U_{(i-j-k_1)}-U_{(i-j)})^2}\right)\\
&- 4\sigma^2\sum_{j=1}^{k_2-k_1}\frac{w_{i,j,2}w_{i,j+k_1,2}}{(U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)})(U_{(i+j+2k_1)}+U_{(i+j+k_1)}-U_{(i-j-2k_1)}-U_{(i-j-k_1)})}\\
&\left(\frac{1}{(U_{(i+j+k_1)}-U_{(i+j)})(U_{(i+j+2k_1)}-U_{(i+j+k_1)})}+\frac{1}{(U_{(i-j-k_1)}-U_{(i-j)})(U_{(i-j-2k_1)}-U_{(i-j-k_1)})}\right)\\
&- 4\sigma^2\sum_{j=1+k_1}^{k_2}\frac{w_{i,j,2}w_{i,j-k_1,2}}{(U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)})(U_{(i+j)}+U_{(i+j-k_1)}-U_{(i-j)}-U_{(i-j+k_1)})}\\
&\left(\frac{1}{(U_{(i+j+k_1)}-U_{(i+j)})(U_{(i+j)}-U_{(i+j-k_1)})}+\frac{1}{(U_{(i-j-k_1)}-U_{(i-j)})(U_{(i-j)}-U_{(i-j+k_1)})}\right)\\
&\leq 4\sigma^2\sum_{j=1}^{k_2}\frac{w_{i,j,2}^2}{(U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)})^2}\left(\frac{2}{(U_{(i+j+k_1)}-U_{(i+j)})^2}+\frac{2}{(U_{(i-j-k_1)}-U_{(i-j)})^2}\right)\\
&= \frac{4(n+1)^4\sigma^2}{k_1^2\sum_{j=1}^{k_2}(2j+k_1)^2}\{1+o_p(1)\}
\end{aligned}$$

provided that $k_1 \to \infty$ and $k_2 \to \infty$ as $n \to \infty$. Both results hold uniformly for $\sum_{j=1}^{2} k_j + 1 \leq i \leq n - \sum_{j=1}^{2} k_j$. In line 4 and 5, there are four covariance terms. $\mathbf{Cov}\big[Y_{i+j+k_1} - Y_{i+j}, Y_{i+l+k_1} - Y_{i+l}\big] = \mathbf{Cov}\big[Y_{i+j+k_1}, Y_{i+l+k_1}\big] - \mathbf{Cov}\big[Y_{i+j}, Y_{i+l+k_1}\big] - \mathbf{Cov}\big[Y_{i+j+k_1}, Y_{i+l}\big] + \mathbf{Cov}\big[Y_{i+j}, Y_{i+l}\big]$. When $j = l$, the first and the fourth covariance are not zero, when $j = l+k_1$ the second covariance is not zero, and when $j + k_1 = l$, the third covariance is not zero. The other three covariance terms can be obtained in a similar way.

3.6.11 Appendix K. Proof of Corollary 4

For $k_1 \to \infty$ and $k_2 \to \infty$ as $n \to \infty$ and from Theorem 3

$$\begin{aligned}
\left|\text{bias}\left[\hat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\right]\right| &\leq \frac{\sup_{u\in[0,1]}|r^{(3)}(u)|}{n+1}\frac{2\sum_{j=1}^{k_2}j^3+3k_1\sum_{j=1}^{k_2}j^2+\frac{5}{3}k_1^2\sum_{j=1}^{k_2}j+\frac{1}{3}k_1^3k_2}{4\sum_{j=1}^{k_2}j^2+k_1^2k_2+4k_1\sum_{j=1}^{k_2}j}\{1+o_p(1)\} \\
&= O_p\left(\max\left\{\frac{k_1}{n},\frac{k_2}{n}\right\}\right) \\
\mathbf{Var}\left[\hat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\right] &\leq \frac{4(n+1)^4\sigma^2}{k_1^2\sum_{j=1}^{k_2}(2j+k_1)^2}\{1+o_p(1)\} \\
&= O_p\left(\max\left\{\frac{n^4}{k_1^2k_2^3},\frac{n^4}{k_1^4k_2}\right\}\right)
\end{aligned}$$

Under the conditions $k_1 \to \infty$ and $k_2 \to \infty$ as $n \to \infty$ such that $n^{-1}k_1 \to 0$, $n^{-1}k_2 \to 0$, $n^4k_1^{-2}k_2^{-3} \to 0$ and $n^4k_1^{-4}k_2^{-1} \to 0$, Theorem 3 states that the conditional bias and conditional variance go to zero. Consequently, we have that

$$\lim_{n\to\infty}\text{MSE}\left[\hat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\right]=\lim_{n\to\infty}\left(\text{bias}^2\left[\hat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\right]+\mathbf{Var}\left[\hat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\right]\right)=0.$$

According to Chebyshev's inequality the proof is complete.

3.6.12 Appendix L. Proof of Corollary 5

From the bias-variance decomposition of the mean squared error (MSE), it follows that

$$\begin{aligned}
\text{MSE}\left[\hat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\right] &\leq \left(\frac{\mathcal{B}_2}{n+1}\frac{2\sum_{j=1}^{k_2}j^3+3k_1\sum_{j=1}^{k_2}j^2+\frac{5}{3}k_1^2\sum_{j=1}^{k_2}j+\frac{1}{3}k_1^3k_2}{4\sum_{j=1}^{k_2}j^2+k_1^2k_2+4k_1\sum_{j=1}^{k_2}j}\right)^2\{1+o_p(1)\} \\
&+ \frac{4(n+1)^4\sigma^2}{k_1^2\sum_{j=1}^{k_2}(2j+k_1)^2}\{1+o_p(1)\}
\end{aligned}$$

with $\mathcal{B}_2=\sup_{u\in[0,1]}|r^{(3)}(u)|$. Since $U\sim\mathcal{U}(0,1)$, the mean integrated squared error (MISE) which measures the average global error is

$$\begin{aligned}
\text{MISE}\left[\hat{Y}^{(2)}|\widetilde{\mathbb{U}}\right] &= \mathbf{E}\int_0^1\left[\hat{Y}^{(2)}(U)-r^{(2)}(U)|\mathbb{U}\right]^2dU \\
&\leq \left(\frac{\mathcal{B}_2}{n+1}\frac{2\sum_{j=1}^{k_2}j^3+3k_1\sum_{j=1}^{k_2}j^2+\frac{5}{3}k_1^2\sum_{j=1}^{k_2}j+\frac{1}{3}k_1^3k_2}{4\sum_{j=1}^{k_2}j^2+k_1^2k_2+4k_1\sum_{j=1}^{k_2}j}\right)^2\{1+o_p(1)\} \\
&+ \frac{4(n+1)^4\sigma^2}{k_1^2\sum_{j=1}^{k_2}(2j+k_1)^2}\{1+o_p(1)\}
\end{aligned}$$

where $\hat{Y}^{(2)}(U)$ represents the second order noisy derivative estimator at the design point U. Denote the asymptotic conditional MISE (AMISE) by

$$\begin{aligned}\text{AMISE}\big[\hat{Y}^{(2)}|\widetilde{\mathbb{U}}\big] &\leq \left(\frac{B_2}{n+1}\frac{2\sum_{j=1}^{k_2}j^3+3k_1\sum_{j=1}^{k_2}j^2+\frac{5}{3}k_1^2\sum_{j=1}^{k_2}j+\frac{1}{3}k_1^3k_2}{4\sum_{j=1}^{k_2}j^2+k_1^2k_2+4k_1\sum_{j=1}^{k_2}j}\right)^2 \\ &+ \frac{4(n+1)^4\sigma^2}{k_1^2\sum_{j=1}^{k_2}(2j+k_1)^2}.\end{aligned}$$

3.6.13 Appendix M. Proof of Exact Bias for the Second Order Derivative

Assume the fourth order derivative of r exist on [0, 1]; using Lemma 1 and weights in Equation (3.16), the exact bias of (3.15) is

$$\begin{aligned}&\text{bias}\big[\hat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\big] = 2\sum_{j=1}^{k_2}w_{i,j,2}\frac{\left(\frac{r(U_{(i+j+k_1)})-r(U_{(i+j)})}{U_{(i+j+k_1)}-U_{(i+j)}}-\frac{r(U_{(i-j-k_1)})-r(U_{(i-j)})}{U_{(i-j-k_1)}-U_{(i-j)}}\right)}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}}-r^{(2)}(U_{(i)}) \\ &= 2\sum_{j=1}^{k_2}w_{i,j,2}\Bigg\{\frac{\left(r'(U_{(i+j)})-r'(U_{(i-j)})+\frac{1}{2}r^{(2)}(U_{(i+j)})(U_{(i+j+k_1)}-U_{(i+j)})\right)}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}} \\ &+\frac{-\frac{1}{2}r^{(2)}(U_{(i-j)})(U_{(i-j-k_1)}-U_{(i-j)})+\frac{1}{6}r^{(3)}(U_{(i+j)})(U_{(i+j+k_1)}-U_{(i+j)})^2}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}} \\ &+\frac{-\frac{1}{6}r^{(3)}(U_{(i-j)})(U_{(i-j-k_1)}-U_{(i-j)})^2+O_p(\frac{k_1^3}{n^3})}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}}\Bigg\}-r^{(2)}(U_{(i)}) \\ &= 2\sum_{j=1}^{k_2}w_{i,j,2}\Bigg\{\frac{\frac{1}{2}r^{(3)}(U_{(i)})[(U_{(i+j)}-U_{(i)})^2-(U_{(i-j)}-U_{(i)})^2]}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}} \\ &+\frac{\frac{1}{2}r^{(3)}(U_{(i)})[(U_{(i+j)}-U_{(i)})(U_{(i+j+k_1)}-U_{(i+j)})-(U_{(i-j)}-U_{(i)})(U_{(i-j-k_1)}-U_{(i-j)})]}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}} \\ &+\frac{\frac{1}{6}r^{(3)}(U_{(i)})[(U_{(i+j+k_1)}-U_{(i+j)})^2-(U_{(i-j-k_1)}-U_{(i-j)})^2]+O_p(\frac{k_1^3}{n^3})+O_p(\frac{j^3}{n^3})}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}}\Bigg\} \\ &= 2\sum_{j=1}^{k_2}w_{i,j,2}\frac{O_p(\frac{j^{\frac{3}{2}}}{n^2})+O_p(\frac{k_1^{\frac{3}{2}}}{n^2})+O_p(\frac{k_1^3}{n^3})+O_p(\frac{j^3}{n^3})}{\frac{4j+2k_1}{n+1}\left\{1+O_p(\max\{j^{-1/2},k_1^{-1/2}\})\right\}} \\ &= O_p\left(\max\left\{\frac{k_1^{\frac{1}{2}}}{n},\frac{k_2^{\frac{1}{2}}}{n},\frac{k_1^2}{n^2},\frac{k_2^2}{n^2}\right\}\right)\end{aligned}$$

3.6.14 Appendix N. Proof of Theorem 4

The proof is analogous to the proof of Theorem 2 in Appendix 3.6.8. Denote $k' = k_1+k_2$, following the proof of Theorem 3.1 in Fan and Gijbels (1996) and based on Theorem 3, for

$k_1 \to \infty$ and $k_2 \to \infty$ as $n \to \infty$ we have

$$\boldsymbol{\epsilon}_1^T \, \mathbf{S}_n^{-1} \, {\mathbf{U}_\mathrm{u}}^T \, \mathbf{W}_\mathrm{u} \begin{bmatrix} \mathrm{bias}[\hat{Y}^{(2)}_{k'+1} | \widetilde{\mathbb{U}}] \\ \vdots \\ \mathrm{bias}[\hat{Y}^{(2)}_{n-k'} | \widetilde{\mathbb{U}}] \end{bmatrix}$$
$$\leq \frac{\mathcal{B}_2}{n+1} \frac{2\sum_{j=1}^{k_2} j^3 + 3k_1 \sum_{j=1}^{k_2} j^2 + \frac{5}{3} k_1^2 \sum_{j=1}^{k_2} j + \frac{1}{3} k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1 \sum_{j=1}^{k_2} j} \boldsymbol{\epsilon}_1^T \, \mathbf{S}^{-1} \, \tilde{c}_p \{1 + o_p(1)\}$$

and for p odd,

$$\begin{aligned} \boldsymbol{\epsilon}_1^T \, \mathbf{S}_n^{-1} \, {\mathbf{U}_\mathrm{u}}^T \, \mathbf{W}_\mathrm{u} \begin{bmatrix} r^{(2)}(U_{(k'+1)}) \\ \vdots \\ r^{(2)}(U_{(n-k')}) \end{bmatrix} - r^{(2)}(u_0) &= \boldsymbol{\epsilon}_1^T \, \mathbf{S}^{-1} \, c_p \beta_{p+1} h^{p+1} + o_p(h^{p+1}) \\ &= \boldsymbol{\epsilon}_1^T \, \mathbf{S}^{-1} \, \frac{c_p}{(p+1)!} r^{(p+3)}(u_0) h^{p+1} + o_p(h^{p+1}). \end{aligned}$$

Combining the two above expressions yields

$$\begin{aligned} \mathrm{bias}\big[\hat{r}^{(2)}(u_0) | \widetilde{\mathbb{U}}\big] &= \mathbf{E}\big[\hat{r}^{(2)}(u_0) | \widetilde{\mathbb{U}}\big] - r^{(2)}(u_0) \\ &= \boldsymbol{\epsilon}_1^T \, \mathbf{S}_n^{-1} \, {\mathbf{U}_\mathrm{u}}^T \, \mathbf{W}_\mathrm{u} \, \mathbf{E}[\hat{\mathbf{Y}}^{(2)} | \widetilde{\mathbb{U}}] - r^{(2)}(u_0) \\ &= \boldsymbol{\epsilon}_1^T \, \mathbf{S}_n^{-1} \, {\mathbf{U}_\mathrm{u}}^T \, \mathbf{W}_\mathrm{u} \left(\begin{bmatrix} r^{(2)}(U_{(k'+1)}) \\ \vdots \\ r^{(2)}(U_{(n-k')}) \end{bmatrix} + \begin{bmatrix} \mathrm{bias}[\hat{Y}^{(2)}_{k'+1} | \mathbb{U}] \\ \vdots \\ \mathrm{bias}[\hat{Y}^{(2)}_{n-k'} | \mathbb{U}] \end{bmatrix} \right) - r^{(2)}(u_0) \\ &\leq \boldsymbol{\epsilon}_1^T \, \mathbf{S}^{-1} \Bigg[\tilde{c}_p \frac{\mathcal{B}_2}{n+1} \frac{2\sum_{j=1}^{k_2} j^3 + 3k_1 \sum_{j=1}^{k_2} j^2 + \frac{5}{3} k_1^2 \sum_{j=1}^{k_2} j + \frac{1}{3} k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1 \sum_{j=1}^{k_2} j} \\ &+ \frac{c_p}{(p+1)!} r^{(p+3)}(u_0) h^{p+1} \Bigg] \{1 + o_p(1)\}. \end{aligned}$$

Plugging (3.33) and (3.34) into this conditional bias gives the second term of the conditional bias of Theorem 4.

According to Theorem 1 in De Brabanter et al. (2018)

$$\begin{aligned} \mathbf{Var}\big[\hat{r}^{(2)}(u_0) | \widetilde{\mathbb{U}}\big] &= \boldsymbol{\epsilon}_1^T \, \mathbf{S}_n^{-1} ({\mathbf{U}_\mathrm{u}}^T \, \mathbf{W}_\mathrm{u} \, \mathbf{Var}[\hat{\mathbf{Y}}^{(2)} | \widetilde{\mathbb{U}}] \, \mathbf{W}_\mathrm{u} \, \mathbf{U}_\mathrm{u}) \, \mathbf{S}_n^{-1} \, \boldsymbol{\epsilon}_1 \\ &\leq \frac{4(n+1)^4 \sigma_e^2}{k_1^2 \sum_{j=1}^{k_2} (2j + k_1)^2} \frac{1 + f(u_0) \rho_c'}{h(n - 2k') f(u_0)} \boldsymbol{\epsilon}_1^T \, \mathbf{S}^{-1} \, \mathbf{S}^* \, \mathbf{S}^{-1} \, \boldsymbol{\epsilon}_1 \{1 + o_p(1)\} \end{aligned}$$

with $\lim_{n\to\infty} n\int \rho_n(x)dx = \rho_c'$ and $\mathbf{S}^* = (\nu_{i+j})_{0\leq i,j\leq p}$ with $\nu_j = \int u^j K^2(u)du$. Plugging (3.33) and (3.34) into this conditional variance gives the second term of the conditional variance of Theorem 4.

3.7 References

Casella, G. and Berger, R. L. (2002). *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.

Charnigo, R., Hall, B., and Srinivasan, C. (2011). A generalized c_p criterion for derivative estimation. *Technometrics*, 53(3):238–253.

Chaudhuri, P. and Marron, J. S. (1999). Sizer for exploration of structures in curves. *Journal of the American Statistical Association*, 94(447):807–823.

Dai, W., Tong, T., and Genton, M. G. (2016). Optimal estimation of derivatives in nonparametric regression. *The Journal of Machine Learning Research*, 17(164):1–25.

David, H. A. and Nagaraja, H. N. (2003). *Order statistics, 3rd Edition*. Wiley Online Library.

De Brabanter, K., Cao, F., Gijbels, I., and Opsomer, J. (2018). Local polynomial regression with correlated errors in random design and unknown correlation structure. *Biometrika*, 105(3):681–690.

De Brabanter, K., De Brabanter, J., De Moor, B., and Gijbels, I. (2013). Derivative estimation with local polynomial fitting. *The Journal of Machine Learning Research*, 14(1):281–301.

Delecroix, M. and Rosa, A. (1996). Nonparametric estimation of a regression function and its derivatives under an ergodic hypothesis. *Journal of Nonparametric Statistics*, 6(4):367–382.

Duong, T. (2018). Kernel smoothing. https://cran.r-project.org/web/packages/ks/index.html.

Eubank, R. L. and Speckman, P. L. (1993). Confidence bands in nonparametric regression. *Journal of the American Statistical Association*, 88(424):1287–1301.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, volume 66. CRC Press.

Gijbels, I. and Goderniaux, A.-C. (2005). Data-driven discontinuity detection in derivatives of a regression function. *Communications in Statistics-Theory and Methods*, 33(4):851–871.

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2006). *A distribution-free theory of nonparametric regression*. Springer.

Hall, P., Kay, J., and Titterinton, D. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, 77(3):521–528.

Härdle, W. (1990). *Applied nonparametric regression*. Cambridge university press.

Hassibi, B. and Stork, D. G. (1993). Second order derivatives for network pruning: Optimal brain surgeon. In *Advances in neural information processing systems*, pages 164–171.

Iserles, A. (2009). *A first course in the numerical analysis of differential equations*. Number 44. Cambridge university press.

Liu, Y. and De Brabanter, K. (2018). Derivative estimation in random design. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 3449–3458. Curran Associates, Inc.

Müller, H.-G. (2012). *Nonparametric regression analysis of longitudinal data*, volume 46. Springer.

Müller, H.-G., Stadtmüller, U., and Schmitt, T. (1987). Bandwidth choice and confidence intervals for derivatives of noisy data. *Biometrika*, 74(4):743–749.

Ojeda Cabrera, J. L. (2012). locpol: Kernel local polynomial regression. https://cran.r-project.org/web/packages/locpol/index.html.

Opsomer, J., Wang, Y., and Yang, Y. (2001). Nonparametric regression with correlated errors. *Statistical Science*, pages 134–153.

Park, C. and Kang, K.-H. (2008). Sizer analysis for the comparison of regression curves. *Computational Statistics & Data Analysis*, 52(8):3954–3970.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076.

Ramsay, J. (1998). derivative estimation. *StatLib -S-News*.

Ramsay, J. O. and Silverman, B. W. (2007). *Applied functional data analysis: methods and case studies*. Springer.

Ramsey, J. and Ripley, B. (2017). pspline: Penalized smoothing splines. https://cran.r-project.org/web/packages/pspline/index.html.

Rice, J. A. (1986). Bandwidth choice for differentiation. *Journal of Multivariate Analysis*, 19(2):251–264.

Rondonotti, V., Marron, J., and Park, C. (2007). Sizer for time series: a new approach to the analysis of trends. *Electronic Journal of Statistics*, 1:268–289.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, pages 832–837.

Stone, C. J. (1985). Additive regression and other nonparametric models. *The annals of Statistics*, pages 689–705.

Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation*. Springer.

Wahba, G. and Wang, Y. (1990). When is the optimal regularization parameter insensitive to the choice of the loss function? *Communications in Statistics-Theory and Methods*, 19(5):1685–1700.

Wang, W. and Lin, L. (2015). Derivative estimation based on difference sequence via locally weighted least squares regression. *Journal of Machine Learning Research*, 16:2617–2641.

Xia, Y. (1998). Bias-corrected confidence bands in nonparametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(4):797–811.

Zhou, S. and Wolfe, D. A. (2000). On derivative estimation in spline regression. *Statistica Sinica*, pages 93–108.

CHAPTER 4. AUTOMATIC CLASSIFICATION OF BLOODSTAIN PATTERNS CAUSED BY GUNSHOT AND BLUNT IMPACT AT VARIOUS DISTANCES

A Paper submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence

Yu Liu, Daniel Attinger, Kris De Brabanter

Abstract

Bloodstain pattern analysis (BPA) plays an important role in forensics towards crime scene analysis and reconstruction. One important question is whether the blood is spattered via gunshot or blunt impact such as beatings. Currently there is no general method to perform this classification. This paper proposes an automated framework to classify the bloodstain spatters caused by either gunshot or blunt impact, based on machine learning. We analyze 94 blood spatters which are being disseminated as free public data sets for research purposes and design features relevant to distinguishing between the gunshot and blunt impact spatters. The study also shows how the distance between the target surface collecting the stains and the blood source influences the bloodstain pattern. Finally we obtain the accuracy of the proposed classification model for different distances ranges. Based on our data set, the proposed model achieves 98.81% accuracy in classifying spatters at distances within 30cm, 93.20% accuracy at distances within 60cm and 85.96% accuracy at distances within 120cm.

4.1 Introduction

A bloodstain pattern is a collection of blood stains. Stains are marks that can be observed on a solid surface called a target. A blood spatter is a bloodstain pattern produced by drops of blood from a blood source which traveled through the air before reaching a target surface, such as a wall or the cardstock used in this study. Bloodstain pattern analysis (BPA) has been used in US criminal courts since the second half of last century, with pioneering contributions of Kirk (1955); Kirk and Kirk (1953); MacDonell and Bialousz (1971). BPA interprets the bloodstain patterns in a crime scene in order to provide evidence to support the crime scene reconstruction MacDonell and Bialousz (1971); Karger et al. (2008). Typically, BPA uses principles of physics, statistics, biology and mathematics to answer questions related to the crime scene. Example questions are: what is the mechanism causing the bloodstain patterns and where is the origin of the blood source. Regarding the latter question, established methods based on the assumption of straight trajectories are widely used to predict the region of origin of a blood spatter Bevel and Gardner (2008); Knock and Davison (2007). de Bruin et al. (2011) proposed a method to improve the determination of the blood source location and Hakim and Liscio (2015) discussed the application of laser scanning technology. Varney and Gittes (2011) introduced a plot-based method to locate the blood source. Further, Camana (2013) used a probabilistic approach to identify the horizontal projection of the location of the blood source.

While there have been multiple academic efforts in estimating the location of the blood source, less attention has been given to assessing the atomization process which atomizes blood into multiple drops and causes the blood spatter, such as beating and shooting discussed here. An important question is whether the bloodstain was generated by a bullet or blunt impact. Studying the bloodstain patterns becomes important to support other evidence such as potential weapons or bullet marks found on the crime scene. MacDonell and Bialousz (1971) proposed to classify spatters between high velocity spatters, usually generated from gunshot and medium velocity blood spatters, i.e. generated from blunt impacts

of club, axe, hammer, fist and brick. They used stain size as a criterion for that classification. They observed in their experiment that "many" stains in a medium velocity spatter exhibit 1/8 inch diameter or smaller, while "essentially all" stains in a high velocity spatter have diameters smaller than 1/8 inch. James et al. (2005) classified medium-velocity and high-velocity impact spatters according to the correlation between the velocity of the force and the size of resulting bloodstains. They stated that :"medium-velocity impact spatters are bloodstains created when the source of blood is subjected to a force with a velocity in the range of 5 to 25 ft/sec ... the diameters of the resulting stains are in the size range of 1 to 3mm, though smaller and larger stains may be present ...", while "the high velocity impact are created when the source of blood is subjected to a force with velocity of greater than 100 ft/sec ...the diameters of the spatters are predominately less than 1mm, although smaller and larger stains are often observed ...".

There is currently no widely accepted method to classify between gunshot and blunt patterns, and the forensic community has abandoned the above distinction between medium-velocity and high-velocity impact spatters Bevel and Gardner (2008). The first reason is that it is difficult to differentiate between the gunshot and blunt patterns. According to MacDonell and Bialousz (1971), both gunshot and blunt spatters exhibit a large number of small stains under 1/8 inch. However, their classification criterion is not quantitative such as "many" and "essentially all under 1/8 inch diameter" indicating that the criterion of percentage of small stains seems to be inconclusive due to the similarity of two patterns. The second problem is that classification by BPA analysts is viewed as subjective and vulnerable to contextual bias Laber et al. (2014). For instance, the phrase "predominately less" used by James et al. (2005) depends on the explanation of an individual BPA analyst. Taylor et al. (2016) assessed the reliability of classification decisions in bloodstain pattern analysis. In that study, well-trained and experienced analysts had a classification accuracy of 69.5%. A third problem is that multiple factors also affect bloodstain spatters, interact with the atomization process and are poorly considered. One factor is the horizontal distance from

blood to target, called hereafter the BT distance, i.e. the horizontal distance between the blood source and the target surface where blood stains reached. For gunshot spatters, if the BT distance is over 3 or 4 feet, the average size of bloodstains will increase because small drops cannot travel as far as larger ones. The reason for this observation stems from basic physics and experience: small snowballs do not travel as far as larger ones, because drag affects the inertia of smaller objects more than that of larger ones. Other factors include the speed and shape of the bullet, which effect is not yet completely understood Comiskey et al. (2018, 2017b,a). BPA analysts are thus seeking for less ambiguous and more objective methods in forensic analysis, and artificial intelligence is certainly a possible way to assist the analyst in classification Arthur et al. (2015).

With the development of digital image analysis, a large number of pattern recognition and machine learning methods have found their way into the field of forensic science. De Chazal et al. (2005) proposed a system that automatically sorts a database of shoeprint pattern categories in response to a reference shoeprint image. Classification of bloodstains patterns caused by different atomization processes becomes more objective by applying digital image-processing methods. Arthur et al. (2017) discussed in details how to automatically obtain measurements and features from a digital image. However they did not explain how to use those features in a classifier. Siu et al. (2017) proposed a framework to distinguish between the forward gunshot and blunt force spatters in a quantitative way. In their work, descriptive statistics were used to select the important features, however they did not show how to classify an unknown spatter and did not mention the performance of their framework.

In this chapter, we develop a stable, quantitative and objective framework to classify the blunt impact and backward gunshot spatters by combining digital image analysis and machine learning methods. Spatters are created under controlled conditions and are being made available to the research community Attinger et al. (2018a,b). Although many methodologies use the video to assist the analysis Siu et al. (2017); James et al. (2005),

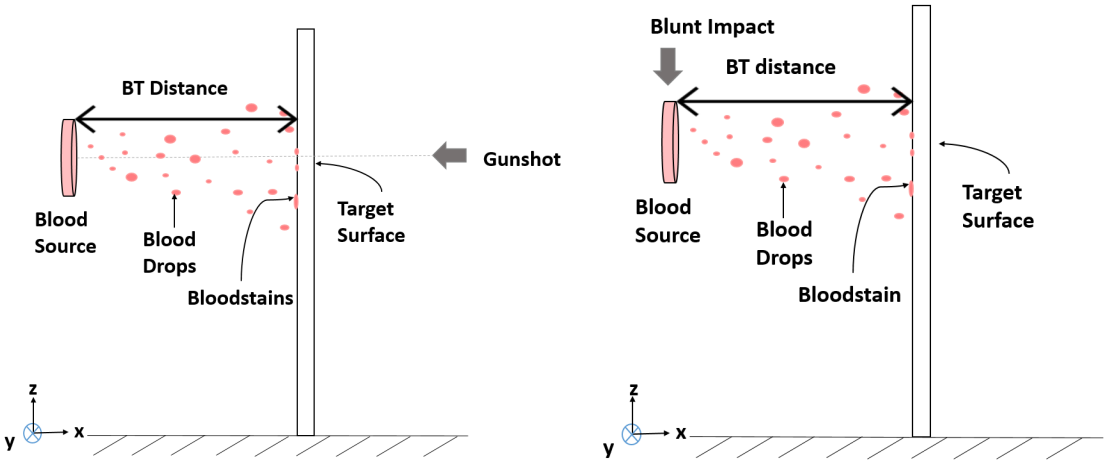
information such as impact velocity is rarely observable in the crime scene. Our framework only requires the static spatter images, which broadens the applicability of this framework. A subset of spatters are used for feature engineering, which is the process to construct features Friedman et al. (2001); James et al. (2013) based on domain knowledge. Later, we randomly split the remaining spatters into a training and a test set. Training sets are used to train the model and the model performance is obtained by comparing the predicted class of each test case with its known true class. We repeat this procedure 1000 times on randomly selected testing and training cases to obtain the average performance of the proposed model on the given spatters Attinger et al. (2018a,b). Meanwhile, we explore the effect of BT distance on features and show the performance of the proposed framework at different BT distances ranges. Our framework achieves 98.81% accuracy in classifying beating spatters and gunshot backspatters at BT distances no larger than 30cm, 93.20% accuracy at BT distances no larger than 60 cm and 85.96% accuracy at BT distances no larger than 120cm.

The major contribution of this chapter consists of three parts. First, we construct different features based on fluid dynamics and discuss these features which are important in classification. Second, we propose a machine learning framework which predicts the cause of an new digital blood spatter, either beating or gunshot. Third, we discuss how BT distance interacts with other features and the influence on the performance of the proposed framework at different BT distances ranges.

4.2 Bloodstain Spatters and Digital Image Preprocessing

The data set used in this paper consists of gunshot backspatters and beating spatters Attinger et al. (2018a,b) which are generated in controlled experiments. Fresh swine blood with anticoagulants is used. Most spatters are generated on a vertical target surface assembled by juxtaposing flat poster board sheets. Since the area of the scanner is smaller

than the target poster, the spatters are scanned in a piecewise manner and are assembled using graphics software (Adobe Photoshop) as high-resolution (600 dpi) jpg files.



(a) Experiment setup for gunshot backspatters.

(b) Experiment setup for beating spatters.

Figure 4.1: Experiment setup for (a) gunshot backspatters and (b)beating spatters. A frame holding the blood source is placed at BT distance away from the target surface. The blood source is either shot or hit to generate the blood spatters. Blood drops from a blood source travel through the air until reaching the target surface, cardstock in this study, resulting in marks called bloodstains on the target cardstock.

Gunshot backspatters are spatters where blood is atomized by a bullet and bloodstains are splashed in a direction opposite to that of the bullet Bevel and Gardner (2008). Fig 4.1a describes the experiment setup to generate the gunshot backspatters. For the spatters investigated in this work, the travel direction of the bullet is perpendicular to both blood source and the target, and a bullet hole is left on the target surface. The bullet hole is hidden from the digitally scanned spatters in the proposed framework because this information is not always available at a crime scene. Two types of gun, handgun and rifle are used to generate the gunshot backspatters, with details in Attinger et al. (2018a,b). Bullet velocity of the rifle is around 1000m/s, and that of the handgun is around 300m/s. Muzzle gases were screened out with a diffuser plate, for the sake of simplicity .

Blunt spatters representative of beating are generated by the impact of a solid object on a blood source(Fig 4.1b). In this data set, two types of beating events were generated. For the so-called hockey puck event, a dowel hits a hockey puck covered with a pool of blood at a impact velocity of 5m/s - 9m/s. For the so-called cylinder event, a sliding cylinder squeezes the blood between two flat surfaces at a impact velocity of 2m/s - 4m/s.

BT distance affects spatters, e.g. the number of stains of a spatter becomes smaller with increasing BT distance. In this paper, we define (atomization) mechanism as the cause of a bloodstain pattern, i.e. handgun, rifle, cylinder and hockey puck. We also define (atomization) process as the general cause of a bloodstain pattern, i.e. gunshot and beating. In order to explore the interaction between BT distance and the mechanism, we generate bloodstains spatters at different BT distances. We only consider 94 single blood source and no muzzle gases spatters within BT distance 120 cm (and including 120cm). The number of replicate experiments under each BT distance and the mechanism pair is shown in Table 4.1.

Table 4.1: Number of replicate experiments

Mechanism	**BT distance(cm)**							
	10	20	30	50	60	90	100	120
Cylinder	0	0	4	0	4	0	0	4
Hockey puck	0	0	9	0	16	0	0	12
Handgun(No Muzzle gases)	3	2	3	4	3	1	2	3
Rifle (No Muzzle gases)	0	0	9	0	7	3	0	5

We apply the image analysis tool *regionprops* in MATLAB (2018) to analyze connected components called spots in spatters. However, some spots are neither bloodstains nor good shaped bloodstains, for example, artificial markers, multiple overlapping bloodstains. We apply the following criteria to find good shaped bloodstains. (1)Due to the scanner resolution, components with area smaller than 0.00021 cm^2 are considered to be noisy. (2) Solidity of a good shaped bloodstain should be larger than 0.75, in which solidity is defined in *regionprops*. (3) Spots with impact angle (arcsine of major axis devided by minor axis of

the fitted ellipse) smaller than $\pi/18$ are discarded because the spot has a high probability to be an artificial mark. (4) Discard spots with Area/FilledArea smaller than 0.95, in which Area and FilledArea are defined in *regionprops*. This criterion removes the obvious overlapping bloodstains.

4.3 Global Features

Due to the comprehensiveness of blood spatter experiments, the number of spatters is not that large Attinger et al. (2018a,b). In order to avoid overfitting Friedman et al. (2001); James et al. (2013), we assign 36 spatters for feature engineering and later obtain the overall performance of the proposed model by repeated random splitting the remaining data set into a training and test set Friedman et al. (2001); James et al. (2013). Features are created based on intuition and knowledge of fluid dynamics. We define the features representing the overall bloodstains in a entire spatter as the global features, the ones representing bloodstains in a specific region of the spatter as local features.

From Table 4.1, at least three spatters were generated by every of mechanism at BT distance 30cm, 60cm, 120cm. We sample 3 images at each experiment setting in {Cylinder, Hockey puck, Handgun, Rifle}$\times\{30cm, 60cm, 120cm\}$ and analyze those 36 spatters to construct features.

4.3.1 Number of bloodstains

The total number of bloodstains in a spatter is affected by the BT distance, the mechanism, etc. Boxplots in Fig 4.2 compare the number of bloodstains of spatters grouped by the BT distance and the mechanism. From the plots, we conclude: (1) For the same mechanism, as the BT distance increases, on average the number of bloodstains in a spatter decreases. This is supported by fluid dynamics since only drops with higher momentum can travel further. (2)On average, gunshot spatters have more bloodstains than beating spatters in the short and medium BT distances ($d \leq$ 60cm), due to the higher velocity of

the bullet which produces high-energy drops compared to blunt impacts. (3) Mechanism influences the number of bloodstains. For example, at distance 30cm, a cylinder impact generates more bloodstains than a hockey puck, even if both cylinder and hockey puck mimicks beating impact. Although gunshot spatters tend to generate more bloodstains, the number of bloodstains is not a good feature for classification (see Fig 4.2).

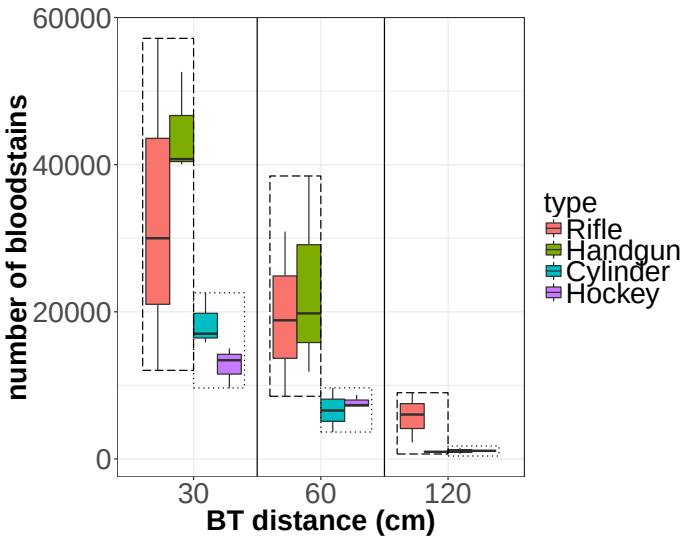


Figure 4.2: Boxplots of number of bloodstains in a spatter against BT distance d and mechanism based on 36 spatters for feature engineering, where $d \in \{30, 60, 120\}$cm. Colors represent different mechanisms. Boxplots belong to either gunshot or beating impact at the same BT distance are grouped by the longdash and dotted rectangle respectively.

4.3.2 Density function of the diameter of the bloodstain

MacDonell and Bialousz (1971), James et al. (2005) and Siu et al. (2017) showed that higher-velocity impacts yield patterns with smaller drop stains. Based on our data, we show that the BT distance along with mechanism impact the density of the diameter of the bloodstain in a spatter.

The diameter of a bloodstain is estimated by diameter $= 2\sqrt{\text{Area}/\pi}$, where Area is the area of the bloodstain in cm^2. We use kernel density estimation(KDE) Parzen (1962); Rosenblatt (1956) to estimate the density of the diameter of the bloodstain in a spatter.

Given $(x_1, x_2, \ldots, x_n)$, KDE at given point x is:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i)$$

where K is gaussian kernel and h is bandwidth. Fig 4.3 compares the KDE of the diameter of the bloodstain of a gunshot backspatter with a beating spatter at BT distance $d = 60$cm, while omitting bloodstains with diameter larger than 3mm. At the same BT distance, a gunshot generates smaller bloodstains than a beating and the dispersion of the diameter of the bloodstain is smaller. Mean and standard deviation of the diameter of the bloodstain are used to capture the difference between gunshot and beating spatters.

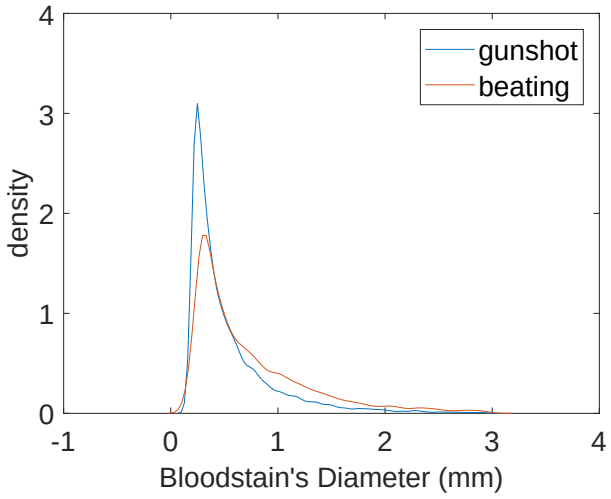
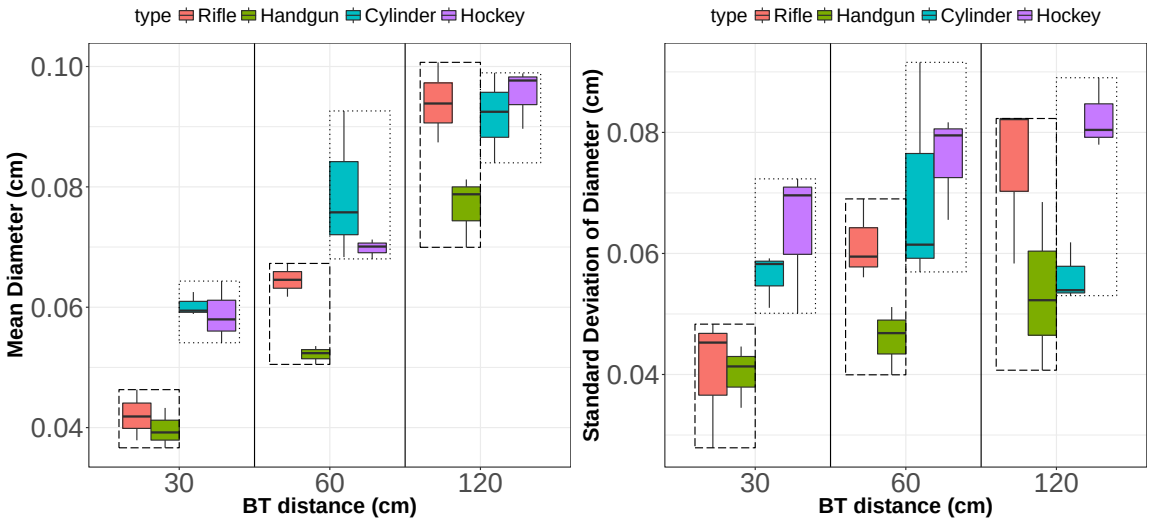


Figure 4.3: Comparison of the KDE of the diameter of the bloodstain ($\leq 3mm$) in a gunshot spatter "GT27" and a beating spatter "C5" at the BT distance $d = 60cm$.

Fig 4.4a and 4.4b show boxplots of sample mean and sample standard deviation of the diameter of the bloodstain in a spatter. (1)Sample mean diameter of the bloodstain classifies the gunshot backspatters and beating spatters at short and medium BT distances($d \leq 60cm$). Gunshots generate smaller bloodstains on average than blunt impact. In addition, sample standard deviation of the diameter of the bloodstain distinguishes between beating and gunshot spatters at short BT distance($30cm$). The dispersion of the diameter of the bloodstain in gunshot backspatters is smaller than that of beating spatters at BT distance

30cm. (2) Those two features are a function of BT distance. The average diameter of the bloodstain in a spatter becomes larger with increasing BT distance, because small blood drops are more affected by air resistance than heavier drops Bevel and Gardner (2008). At large BT distance (120cm), the sample mean bloodstain diameter of the bloodstain cannot differentiate between gunshot and beating spatters. Similarly, the dispersion of the diameter of the bloodstain is increasing for most mechanisms as the BT distance increases.



(a) Mean diameter of the bloodstain(cm)

(b) Standard deviation of the diameter of the bloodstain(cm)

Figure 4.4: Boxplots of (a) sample mean diameter of the bloodstain (cm) (b)sample standard deviation of the diameter of the bloodstain(cm) in a spatter against BT distance and mechanism based on 36 spatters for feature engineering. Longdash and dotted rectangles represent gunshot and beating impact boxplots respectively.

We also incorporate estimations of higher order moments, e.g. skewness and kurtosis, but according to the classifier those features are not significant to differentiate between beating and gunshot spatters.

4.3.3 Fraction of large stains

Fraction of large bloodstains is important in determining the process of a spatter Siu et al. (2017). We further explore this feature by varying the threshold criterion for large

bloodstains and study how BT distance interacts with the spatter mechanism with respect to the fraction of large stains.

For each spatter, we calculate the fraction of bloodstains with area larger than a threshold s, fraction $= \frac{\mathbb{1}[\text{Area}>s]}{n}$, where n is the total number of bloodstains in a spatter and $\mathbb{1}[A]$ is the indicator function of an event A, i.e. if A is true, $\mathbb{1}[A] = 1$ and zero otherwise.

For $s \in \{\pi(\frac{0.3}{2})^2, \pi(\frac{0.5}{2})^2, \pi(\frac{0.8}{2})^2, \pi(\frac{1}{2})^2, \pi(\frac{3}{2})^2\} mm^2$, thresholds $\pi(\frac{0.8}{2})^2 mm^2$ and $\pi(\frac{1}{2})^2 mm^2$ are best at discriminating between gunshot and beating spatters based on 36 spatters used for feature engineering. The fraction of large bloodstains is highly correlated with average size of the bloodstain. The sample Pearson correlation coefficient Pearson (1895) between the mean area of the bloodstain and fraction of bloodstains larger than $\pi(\frac{0.8}{2})^2 mm^2$ is 0.98. The box plots for the fraction of bloodstains larger than $\frac{\pi}{4}0.8^2\ mm^2$ in a spatter are shown in Fig 4.5. This feature distinguishes between the beating and gunshot spatters at BT distance 30cm. Less than 15% of the bloodstains in a gunshot backspatter are larger than $\frac{\pi}{4}0.8^2\ mm^2$, while beating generates more than 15% large bloodstains at BT distance 30cm. As the BT distance increases from 30cm to 120cm, the fraction of large bloodstains increases for all spatters since small bloodstains cannot travel far, resulting in difficulties in classification using this feature.

4.3.4 Vertical difference between large and small bloodstains

In Fig 4.1 we define a three-dimensional coordinate system where the z direction is opposite to the gravity direction and the target cardstock is in the y-z plane at $x = d$ (BT distance). The origin of the two-dimensional coordinate system is at the lower left of the target cardstock. We construct a feature to measure the average distance between large and small bloodstains along the z direction. According to fluid dynamics, air resistance affects the velocity of smaller blood drops more. If the initial velocity is the same for small and large blood drops, small blood drops will eventually hit on the cardstock in a position lower than the large blood drops because gravity has more time to modify their trajectories.

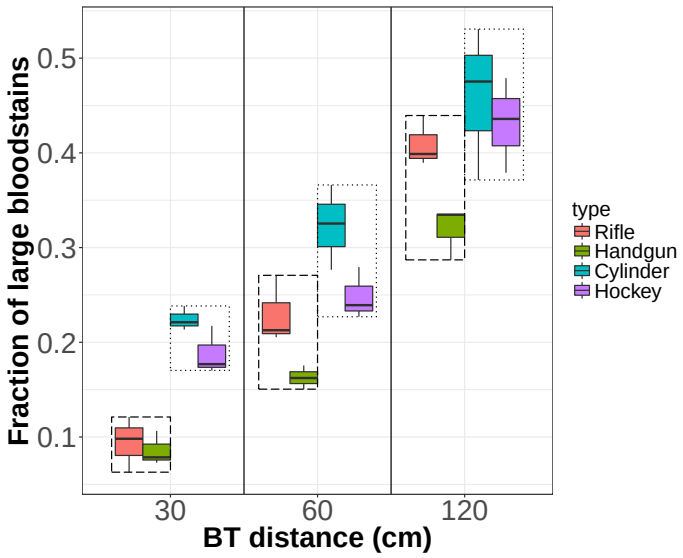


Figure 4.5: Boxplots of fraction of bloodstains larger than $\pi(\frac{0.8}{2})^2 mm^2$ in a spatter against BT distance and mechanism based on 36 spatters for feature engineering. Boxplots belong to either gunshot or beating impact at the same BT distance are grouped by the longdash and dotted rectangle respectively.

Since the blood drops generated by a bullet have higher initial velocity, they are affected by gravity in a shorter period than blood drops generated by blunt impact, possibly leading to different distance between large and small bloodstains along the z direction.

Motivated by this, we rank bloodstains in a spatter by their areas and define the bloodstains whose area falls in the range [12.5th percentile, 37.5th percentile] as small bloodstains and the ones with area within [62.5th percentile, 87.5th percentile] as large bloodstains. Bloodstains with area smaller than 12.5th percentile or larger than 87.5th percentile contain outliers and are removed. The thresholds used to define large and small bloodstains are selected based on trial and error on 36 spatters used for feature engineering. Finally, we construct the following feature delta.z:

$$\begin{aligned} \text{delta.z} \quad &\equiv \quad \frac{1}{\lceil 87.5n \rceil - \lceil 62.5n \rceil + 1} \sum_{i=\lceil 62.5n \rceil}^{\lceil 87.5n \rceil} Z_i \\ &- \quad \frac{1}{\lceil 37.5n \rceil - \lceil 12.5n \rceil + 1} \sum_{i=\lceil 12.5n \rceil}^{\lceil 37.5n \rceil} Z_i, \end{aligned}$$

where Z_i indicates the vertical distance between the i_{th} largest bloodstain to the bottom of the spatter, $\lceil . \rceil$ denotes the ceiling function and n is the number of bloodstains in the spatter. The minuend is the mean vertical position of the large bloodstains, while the subtrahend is that of small bloodstains. From Fig 4.6, the proposed feature differentiates between gunshot and beating spatters at BT distance 60cm. Similar to our earlier discussion, the gunshot spatters have less vertical difference between large and small bloodstains than beating spatters. However, at BT distance 120cm, most bloodstains are large for both gunshot and beating spatters leading to difficulties in differentiating spatters using the proposed feature.

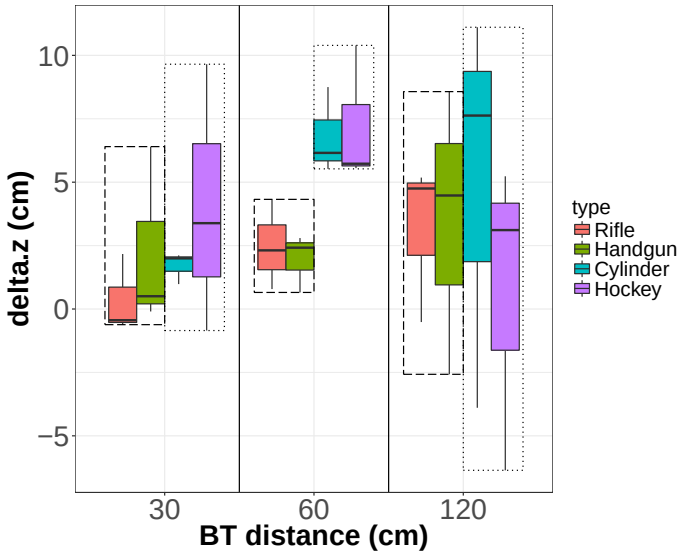


Figure 4.6: Boxplots of delta.z in a spatter against BT distance and mechanism based on 36 spatters for feature engineering. Longdash and dotted rectangles represent gunshot and beating impact boxplots respectively.

4.3.5 Shape of the bloodstain

Every bloodstain is automatically fitted with an ellipse as shown in Fig 4.7.

The shape measurements are:

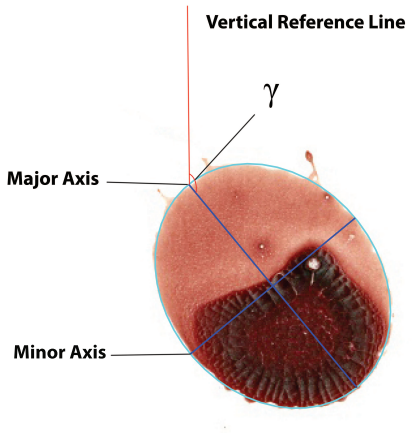


Figure 4.7: Illustration of the fitting procedure with an ellipse (cyan ellipse). Shape measurement γ of a bloodstain based on the fitted ellipse is defined by the angle between the vertical reference line and the major axis.

1 Impact angle α (in degrees):

$$\alpha \equiv \frac{180}{\pi} \arcsin\left(\frac{mAxis}{MAxis}\right)$$

where $mAxis$ and $MAxis$ represent the length of the minor axis and the length of the major axis of the fitted ellipse, see Fig 4.7.

2 Adjusted impact angle α_ϵ: Some bloodstains have long tails, thus we shrink the major axis of fitted ellipse to better fit the part without tail.

$$Adj\_MAxis = MAxis\left(\frac{FilledArea}{EllipsArea}\right)^{tail\_reduction}$$
$$\alpha_\epsilon = \frac{mAxis}{Adj\_MAxis}$$

where $EllipsArea$ is the number of pixels in the fitted ellipse. $FilledArea$ is the number of pixels in a filled image returned from *regionprops* in MATLAB (2018), $tail\_reduction = 3$ and is selected ad hoc.

3 γ: Angle between the vertical reference line (z axis) and the major axis of the fitted ellipse in Fig 4.7. $\gamma = 90° -$ Orientation, where Orientation is obtained from *regionprops* in MATLAB (2018).

4 Eccentricity: The ratio of the distance between the foci of the fitted ellipse and its major axis length, measuring the circularity of the ellipse

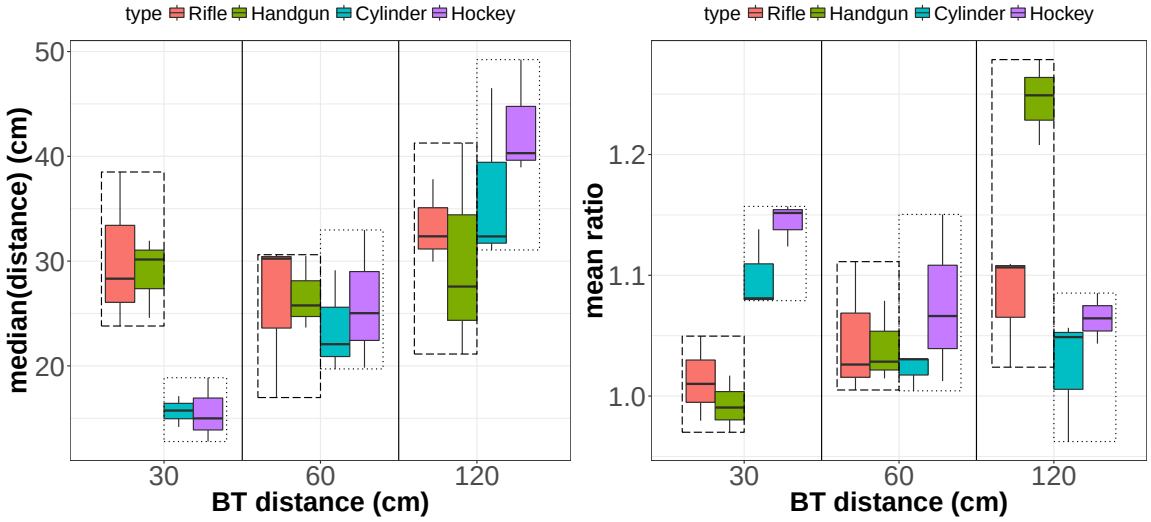
5 Solidity: Proportion of the pixels in the convex hull that are also in the region.

We constructe the sample mean and sample standard deviation of the shape measurements of bloodstains in a spatter as 10 new global features. However, those global features turned out to be less important in classification and we advocate to use shape measurements as local features in Section 4.4.

4.3.6 Distance of bloodstains to the centroid

We construct features which contain the location information of bloodstains in a spatter. Following the coordinate system in Fig 4.1, the target cardstock is in the two-dimensional coordinate system defined by two perpendicular coordinate axes y and z. Given n bloodstains in a spatter, the observed positions for all bloodstains are (Y_i, Z_i) for $i = 1, \dots n$ from the distribution of (Y, Z). We first define the centroid of a spatter as the sample median position of all bloodstains in the spatter as $(\text{median}(Y), \text{median}(Z))$. Given that there could be outliers in the spatter, we choose the median instead of the mean.

Define the random variable *distance* as the Euclidean distance between the bloodstain and the centroid in a spatter. Given n bloodstains in a spatter, the observed Euclidean distances between bloodstains and the centroid are denoted as $distance_i, i = 1, \dots n$. The sample median of these distances is a new feature median(*distance*) which depicts the average spread of bloodstains in a spatter. The boxplots of the proposed feature are shown in Fig 4.8a. At BT distance 30cm, the median of gunshot boxplots are larger than that of beating boxplots, indicating that the spread of bloodstains of gunshot spatters is larger than that of beating spatters and also the spread of bloodstains in a spatter increases as BT distance increases (except gunshot spatters at 30 cm). We use this property to construct local features in Section 4.4.



(a) median(distance) (b) $\frac{\text{mean}(distance)}{\text{median}(distance)}$.

Figure 4.8: Boxplots of (a) median(distance) (b)$\frac{\text{mean}(distance)}{\text{median}(distance)}$ in a spatter against BT distance and mechanism based on 36 spatters for feature engineering. Longdash and dotted rectangles represent gunshot and beating impact boxplots respectively.

To minimize the effect of BT distance on the spread of a spatter, we construct a new random variable $\frac{distance}{\text{median}(distance)}$, which can be interpreted as normalized *distance*. We construct the following two features to reflect the distribution of $\frac{distance}{\text{median}(distance)}$for each spatter: (1) sample mean of $\frac{distance}{\text{median}(distance)}$ (2) sample standard deviation of $\frac{distance}{\text{median}(distance)}$. In Fig 4.8b, $\frac{\text{mean}(distance)}{\text{median}(distance)} > 1$ for beating spatters at BT distance 30cm and it is larger than gunshot spatters, indicating the distribution of distances for a beating spatter is more skewed to right at BT distance 30cm than a gunshot spatter.

4.4 Binning Methodologies and Local Features

All features constructed so far are global features which contains the information for the entire spatter. In this section, we split the whole spatter into small pieces and construct features locally at every piece. Each piece is called a bin.

4.4.1 Various binning methodologies

4.4.1.1 Fixed bins: annulus with fixed width

Concentric circles are circles with a common center, which is the centroid of spatter in this paper. An annulus is a ring-shaped object, a region bounded by two concentric circles. The width of the annulus is $r_0 = R - r$ where R is the radius of the outside circle and r is the radius of the inner circle. Following the convention of Siu et al. (2017), a bloodstain spatter is segregated into 40 annuli with 2.5 cm width that radiate outward from the pattern's centroid. Fig 4.9 illustrates this process conceptually and each annulus is labeled with an index $i = 1, \ldots 40$. The innermost annulus is labeled as bin 1 with the centroid as its center and radius r_0. For annulus with index i, the radius of inner circle is $(i-1)r_0$ and radius of outside cirle is $i \cdot r_0$. We set $r_0 = 2.5cm$ for all spatters.

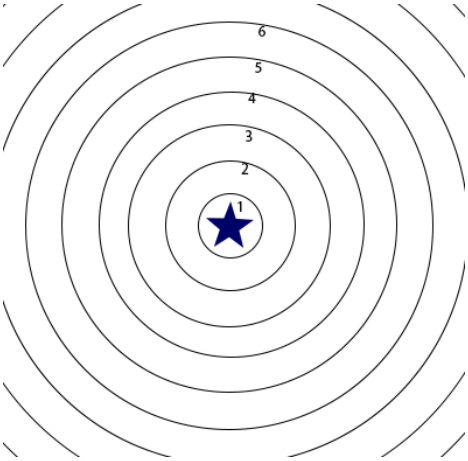


Figure 4.9: Concentric ring bins. The star represents the position of the centroid in the spatter.

4.4.1.2 Adaptive bins: annulus with width proportional to the median distance

Similar to Section 4.4.1.1, a bloodstain spatter is segregated into 40 equidistant annuli that radiate outward from the pattern center as shown in Fig 4.9. However, r_0 =

$\frac{2}{40}$median($distance$), which varies among spatters. The definition of median($distance$) is discussed in Section 4.3.6. The total number of annuli equates to 40 and the constant 2 is chosen since more than 98% of bloodstains fall within the distance of $2 * \text{median}(distance)$ away from the centroid.

Fig 4.10 illustrates why we vary r_0 among spatters. If the gravity and air resistance is ignored, the trajectory of bloodstains is contained in a cone Comiskey et al. (2018, 2017b,a), whose base is rougly circular on the target surface. Moving the blood source towards the target surface, the area of the base progressively becomes smaller. In order to have a fair amount of bloodstains in bin i, $i = 1, \ldots, 40$ for spatters at different BT distances, r_0 is preferred to be a function of the area of the base. The median($distance$) is used to reflect the area of the base. In this way, the bin with index i is comparable among spatters at different BT distances. Fig 4.11 shows an example comparing the two bin definitions. Rifle spatters at BT distance 30 and 120cm are split into 40 bins according to the definition of fixed bins and adaptive bins. The fraction of bloodstains in bin i in spatter j at BT dsitance d is the ratio of number of bloodstains in the ith bin to the total number of bloodstains in spatter j and denoted as $fraction_{ijd}$, where $i = 1, \ldots, 40, j = 1, 2, 3, d = 30cm, 120cm$. The point of errorbar plots at $x = i$ in Fig 4.11 is defined as $\overline{fraction_{i.d}} = \frac{1}{3}\sum_{j=1}^{3} fraction_{ijd}$, which is over 3 replicated spatters($j = 1, 2, 3$) at distance d and bin i. Half length of each error bar is the standard error of the mean over three trial replicates. We connect points with curves to easily observe the curvature information.

Fig 4.11a illustrates that if using fixed bins and at the short BT distance 30cm, almost no bloodstains fall further than bin 22. While for BT distance 120cm, still a fair amount of bloodstains appear in bin 22 to 30. BT distance has a significant impact on the bloodstain pattern. Adaptive binning used in Fig 4.11b provides an alternative way to align bins. The curves at BT distance 30cm and 120cm are expanded such that most bins contain bloodstains. In addition, the curves are smoothed especially for small BT distance, leading

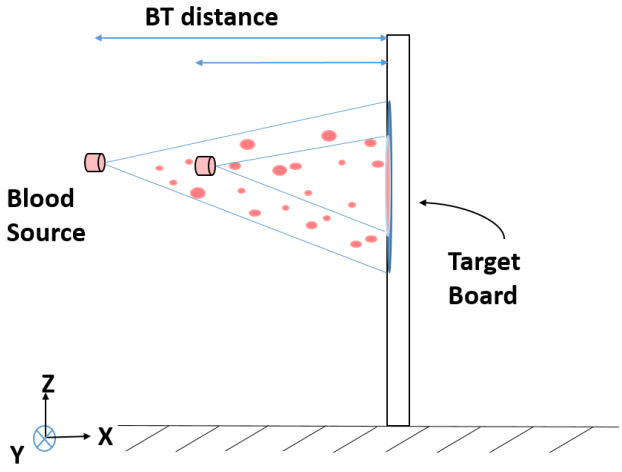
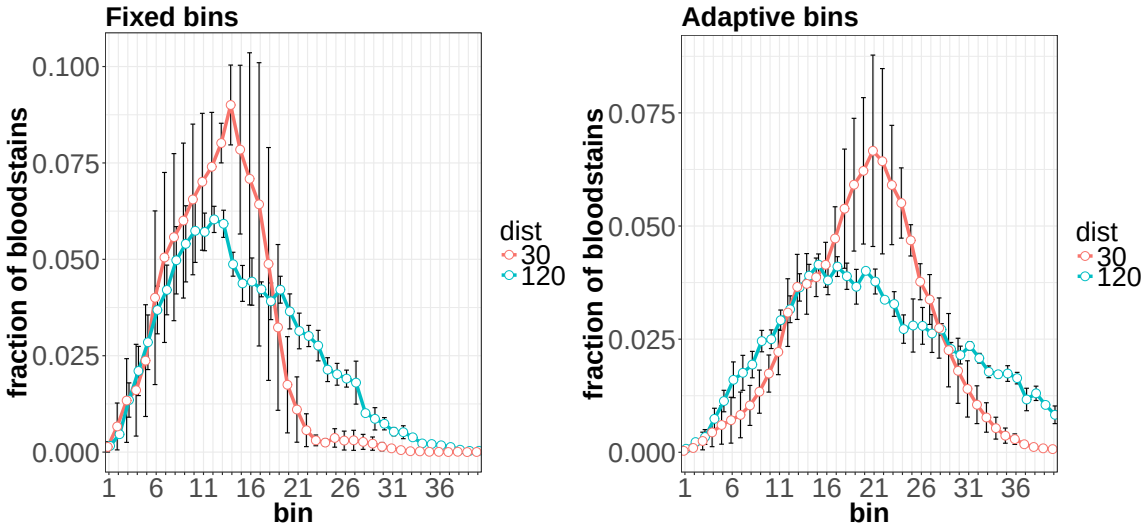


Figure 4.10: Cone shape bloodstains trajectory if gravity and air resistance are ignored. Spatters close to the target cardstock generate spatters with smaller spread. Concentric ring bins.

to the possibility to explore the derivatives over bins between gunshot and beating spatters in Section 4.4.2.3.

4.4.1.3 Rectangular bins

An asymmetric definition of bins is to segregate the spatter into equal size rectangular bins along the vertical direction. Define the random variable Z as the vertical distance between the bloodstain to the bottom of the spatter. Given n bloodstains, the observed data are $Z_i, i = 1, \ldots n$. The width of each rectangle is the same as the width of the spatter and the length of rectangle is $\frac{6}{40} median(|Z - median(Z)|)$, where $median(|Z - median(Z)|)$ is the median absolute deviation of vertical distance(in z direction) between the bloodstain to the centroid. The total number of bins equals 40 and the factor 6 is chosen so that more than 81% of the bloodstains falling within a distance of $3median(|Z - median(Z))$ from the centroid in the vertical direction and are covered by rectangular bins (see Fig 4.12). Both the lower side of rectangle 21 and the top side of rectangle 20 go through the centroid.



(a) Fixed bins

(b) Adaptive bins

Figure 4.11: Error bar plots of the fraction of bloodstains in each bin against the index of the bin grouped by BT distance 30 and 120cm for Rifle spatters in feature engineering data set. Points and half length of the bars in errorbar plots represent the mean and the standard error of the mean over three trial replicates.

4.4.2 Local features

We illustrate the construction of local features with the following example: the fraction of bloodstains with diameter larger than 1mm in each bin.

4.4.2.1 Fraction of large stains in bins

We select the threshold $1mm$ because it is the best at discriminating between gunshot and beating spatters based on feature engineering data set.

(1) Select spatters at a fixed BT distance, e.g. 60cm.

(2) Choose a binning method, e.g. fixed bins with a width of 2.5cm.

(3) For bin $i, i = 1, \ldots 40$, calculate the ratio of bloodstains with diameter larger than 1mm in bin i to the number of bloodstains in bin i, which is ratio_i.

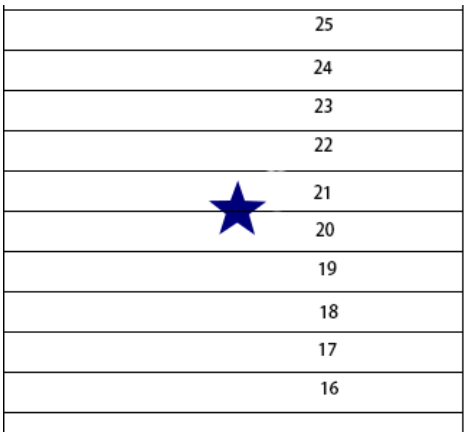


Figure 4.12: Rectangle bins. The star represents the position of centroid in the spatter

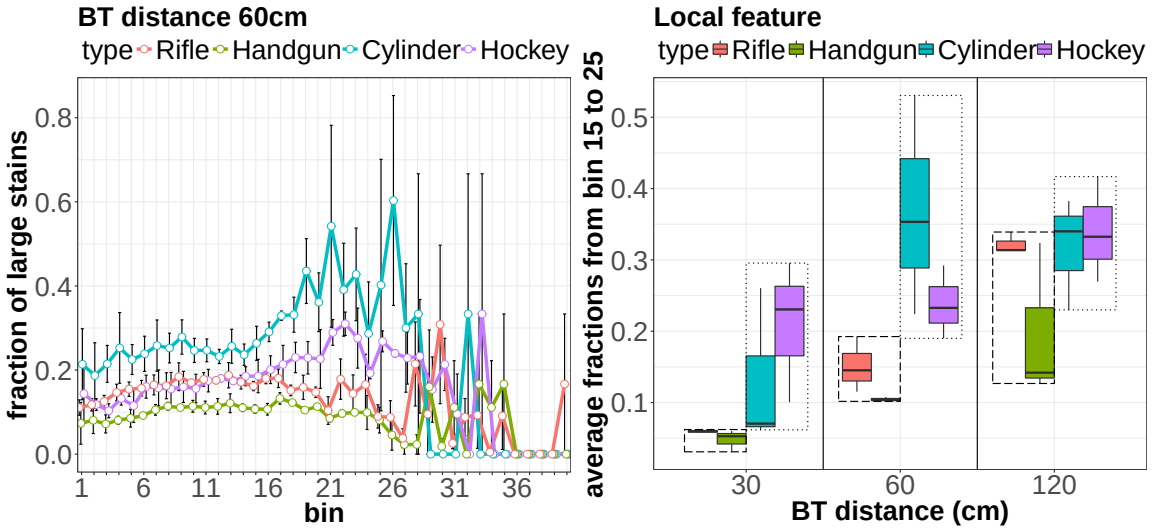
(4) Make the errorbar plots of the fraction of bloodstains with diameter larger than 1mm against the index of the bin over 3 replicated spatters (see Fig 4.13a).

(5) Select s set of bins S in which ratio_i shows a difference between gunshot and beating spatters.

(6) construct the local feature LF by simple average: $\text{LF} = \frac{1}{|S|}\sum_{i\in S}\text{ratio}_i$, where $|S|$ is the cardinality of the set S.

Although it is difficult to distinguish the spatters at BT distances 120cm, for BT distances 30cm and 60cm, in bins with index $S = \{15, 16, \ldots, 25\}$, the fraction of large bloodstains of gunshot spatters is significantly smaller than that of beating spatters. The simple average of the fraction of bloodstains with diameter larger than 1mm from bin 15 to 25 in a spatter is a new feature and its boxplots are shown in Fig 4.13b. The proposed feature distinguishes beating and gunshot spatters at BT distance 30 and 60cm.

Following this procedure, one can construct different features by selecting different combinations of the set of bins, binning method and averaging method.



(a) Errorbarplots.

(b) Boxplots.

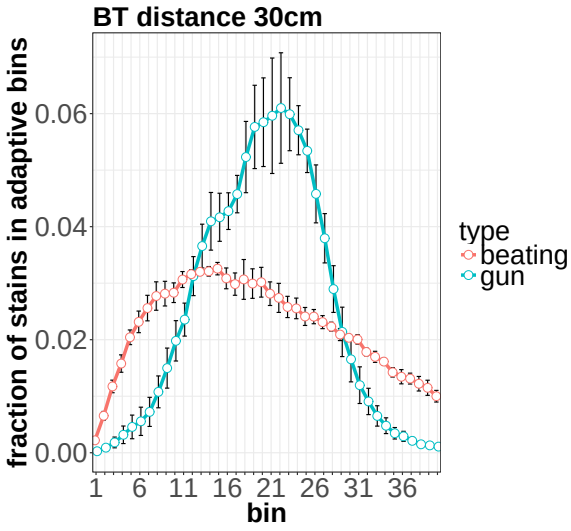
Figure 4.13: (a)Errorbar plots of the fraction of bloodstains with diameter larger than 1mm against the index of the bin for spatters at BT distance 60cm grouped by mechanism. Points and half length of the error bar represent the mean and the standard error of the mean over three trial replicates. (b)Box plots of the average of the fraction of large stains with the diameter larger than 1mm from bin 15 to bin 25 against BT distance grouped by mechanism. Longdash and dotted rectangles represent gunshot and beating impact boxplots respectively. Both (a) and (b) are based on the 36 spatters used for feature engineering.

4.4.2.2 Shape of stains in bins

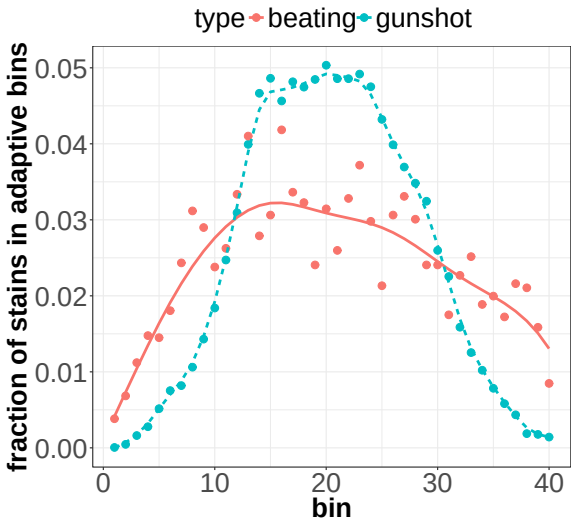
As discussed in Section 4.3.5 though shapes of bloodstains for entire the spatter are not important features in classification, we follow the procedure in Section 4.4.2.1 to construct some useful local features regarding to the shape of bloodstains. For example, one important feature is the mean impact angle α of bloodstains in adaptive bins $23, 24, \ldots 30$

4.4.2.3 Shape of curve over adaptive bins

Fig 4.14a shows the errorbar plots of the fraction of bloodstains in each adaptive bin for spatters at BT distance 30cm grouped by process (gunshot or beating). We plot a curve connecting the points which are the mean fractions over 3 replicated spatters, with red and blue curves representing the beating and gunshot spatters respectively.



(a) Errorbar plots.



(b) Fraction of stains in each adaptive bin and regression curves.

Figure 4.14: (a) Errorbar plots show the fraction of bloodstains in each adaptive bin against the index of the bin for spatters at BT distance 30cm grouped by process based on feature engineering data. Point and half length of the error bar represent the mean and the standard error of the mean over three replicates. (b)Blue points are fractions of bloodstains in adaptive bins for the gunshot spatter "GT45" and Red points are for the beating spatter "C6". Both spatters are generated at BT distance 60cm. Blue dashed line and red solid line are the local linear regression curves for the gunshot and beating spatter respectively.

To smooth the data, we use local polynomial regression Fan and Gijbels (1996). Consider the data $(X_1, Y_1), \ldots, (X_n, Y_n)$ which form an independent and identically distributed (i.i.d.) sample from a population (X, Y). where X_i is the index of the ith adaptive bin, Y_i is the the fraction of bloodstains in ith adaptive bin and n is the total number of bins. The local polynomial regression estimator in an arbitrary point x is given by minimizing the following weighted least squares problem Fan and Gijbels (1996)

$$\min_{\beta_j \in \mathbb{R}} \sum_{i=1}^{n} \left\{ Y_i - \sum_{j=0}^{p} \beta_j \left(X_i - x\right)^j \right\}^2 K_h(X_i - x), \tag{4.1}$$

where β_j are the solutions to the weighted least squares problem. We select the Gaussian kernel and $K_h(u) = \frac{1}{\sqrt{2\pi}h} e^{-\frac{u^2}{2h^2}}$. h is the bandwidth selected from $\{1, 1.006, \ldots, 4\}$. $\hat{m}^{(q)}(x) = q!\hat{\beta}_q$ is an estimator for the q-th order derivative where $q = 0, \ldots, p$. In Fig 4.14b,

red and blue points represent the fraction of bloodstains in each adaptive bin for a beating and gunshot spatter at BT distance 60cm respectively. The fitted regression curves ($q = 0$) are estimated using (4.1) with $p = 1$ for beating and gunshot separately. From both plots in Fig 4.14 it is obvious that the regression function for gunshot is convex over bins $1, \ldots, 10$ and $30, \ldots, 40$ while concave for beating spatters. In addition, the absolute values of the first order derivatives for the gunshot curve are larger than the beating curve over bins 10 to 15 and 25 to 30. We construct the estimated first and second order derivatives at different adaptive bins using local quadratic and cubic regression as new features to measure the shape of curves.

4.5 Modeling

4.5.1 Random forest

Random forests applies bagging to decision tree (Liaw et al., 2002; Breiman, 2001; Cutler et al., 2012).

Algorithm 1 Random Forests for Classification.

For $b = 1$ to B:
(a) Bootstrap sample from the training set.
(b) Fit a classification tree to bootstrapped sample by splitting every terminal node whose size is larger than the minimum size of terminal nodes:
  (a) Randomly sample m features.
  (b) Pick the best split among m and split node into two children nodes.
The predictions for an unseen sample is made by taking the majority votes of classification trees.

4.5.2 Features

A brief description of each feature is shown in Table 4.2, where the last column indicates the section where the feature is introduced and explained in details. Holding out 36 spatters for feature engineering, the feature matrix for the remaining spatters is 58×58, which is composed of 58 features from Table 4.2 for 58 spatters.

Table 4.2: All features used for classification

feature name	**brief description**	**description in Section**
number.stains	number of bloodstains	4.3.1
Diameter.mean	mean diameter of the bloodstain	4.3.2
Diameter.std	standard deviation diameter of the bloodstain	4.3.2
fc	ratio of number of bloodstains with area larger than $\frac{\pi}{4}0.8^2 mm^2$ to the total number of bloodstains	4.3.3
fd	ratio of number of bloodstains with area larger than $\frac{\pi}{4}1^2\ mm^2$ to the total number of bloodstains	4.3.3
delta.z	vertical difference between large and small bloodstains	4.3.4
mean.ratio	$\frac{\text{mean}distance}{\text{median}(distance)}$, where $distance$ is the Euclidean distance between the bloodstain and the centroid.	4.3.6
sd.ratio	sample standard deviation of $\frac{distance}{\text{median}(distance)}$.	4.3.6
fraction1_bin_15_25	average of the fraction of bloodstains with area larger than $\frac{\pi}{4}0.1^2 mm^2$ in fixed bins $15, 16, \ldots, 25$	4.4.2.1
fraction1_bin_35_40	average of the fraction of bloodstains with area larger than $\frac{\pi}{4}0.1^2 mm^2$ in fixed bins $35, 36, \ldots, 40$	4.4.2.1
fraction075_bin_1_17	average of the fraction of bloodstains with area larger than $\frac{\pi}{4}(0.75)^2 mm^2$ in fixed bins $1, 2, \ldots, 17$	4.4.2.1
fraction075_bin_20_21	average of the fraction of bloodstains with area larger than $\frac{\pi}{4}(0.75)^2 mm^2$ in fixed bins $20, 21$	4.4.2.1
fraction075_adp_25_31	average of the fraction of bloodstains with area larger than $\frac{\pi}{4}(0.75)^2 mm^2$ in adaptive bins $25, \ldots, 31$	4.4.2.1
fraction1_adp_27_30	average of the fraction of bloodstains with area larger than $\frac{\pi}{4}(0.1)^2 mm^2$ in adaptive bins $27, \ldots, 30$	4.4.2.1
alpha_adp_23_30	average of mean impact angle of the bloodstain over adaptive bins $23, 24, \ldots, 30$	4.4.2.2, 4.3.5
epsilon_adp_23_30	average of mean adjusted impact angle of the bloodstain over adaptive bins $23, 24, \ldots, 30$	4.4.2.2, 4.3.5
epsilon_rec_27_30	average of mean adjusted impact angle of the bloodstain over rectangle bins $27, 28, \ldots, 30$	4.4.2.2, 4.3.5
derivatives2bin$i$	Second order derivatives of the regression function of the fraction of bloodstains over the adaptive bin at bin $i, i = 1, \ldots, 9, 30, \ldots, 39$	4.4.2.3
derivatives1bin$i$	First order derivatives of the regression function of the fraction of bloodstains over the adaptive bin at bin $i, i = 5, \ldots, 15, 25, \ldots, 35$	4.4.2.3

Holding out 36 spatters used for feature engineering, 58 spatters are left for training and testing. BT distances for spatters are treated as unknown. We use the procedure to obtain the average performance and variability of the model:

(1) Randomly split spatters into a training(75%) and test(25%) set. Values of the label $y \in \{beating, gunshot\}$.

(2) Use random forest (Liaw et al., 2002; Breiman, 2001; Cutler et al., 2012) on the training set and obtain predictions on the test set.

(3) Compare the predicted labels and true label for test set.

(4) Repeat (1)-(3) 1000 times to obtain the average performance and variability of the model.

In step (2), the number of features randomly sampled as candidates at each split of each tree is set to $\lceil\sqrt{58}\rceil$, and the minimum size of the terminal nodes in each tree is 1. The number of trees is set to 5000.

4.5.3 Simulation results

Two metrics are used to evaluate the performance of the classifier Liaw et al. (2002); Breiman (2001):

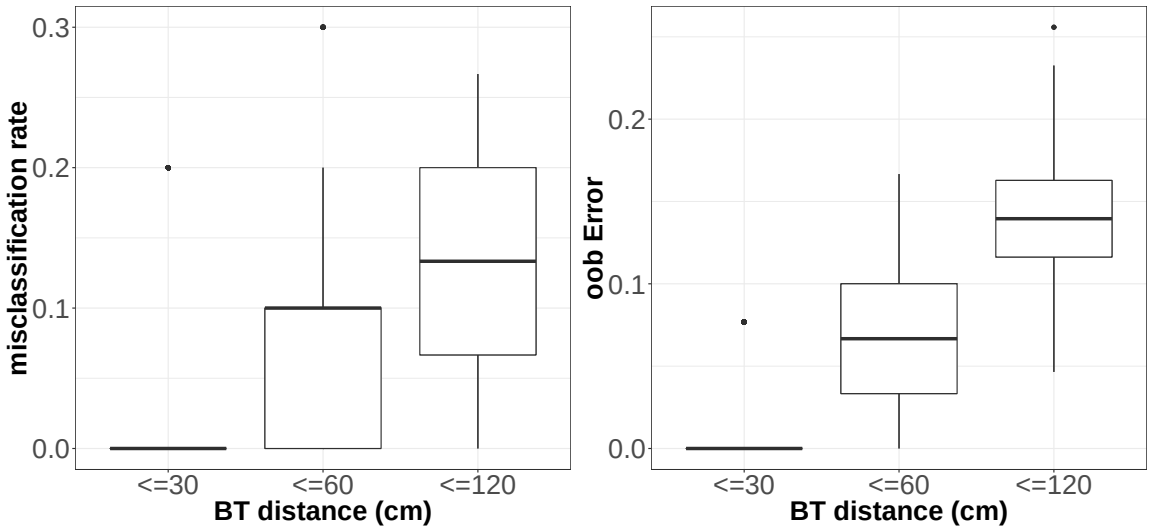
(1) misclassification(MSC) rate:

$$\frac{\mathbb{1}[\text{predicted label} \neq \text{actual label}]}{\text{size of test set}}$$

(2) Out-of-bag(oob) error:

For each tree in the random forest, only 2/3 of the bootstrap sample from the original data is used to construct the tree and the rest is used to estimate the error rate, which is called the Out-of-bag(oob) error. Both MSC rate and oob error are obtained at each iteration.

The performance of the proposed classifier is shown in Fig 4.15 based on the MSC rate and oob error over 1000 replications. In order to study the performance of the proposed classifier at different BT distance ranges among 58 spatters, we perform a Monte Carlo simulation on the spatters with BT distance no larger than 30cm, 60cm and 120cm separately. Table 4.3 shows the total number of spatters used for training(75%) and testing(25%), the mean MSC error, the mean oob error and the mean accuracy, where accuracy $= 1 -$ oob error in each replication. For spatters at short BT distances $d \leq 30$cm, the proposed classifier achieves roughly 98.81% accuracy and for spatters with BT distances $d \leq 60$cm, the proposed classifier achieves 93.20% accuracy on average. For spatters with BT distances $d \leq 120$cm, the proposed classifier achieves 85.96% accuracy on average. From the discussion in Section 4.3 and Section 4.4, many features, "Diameter.mean", "Diameter.std", "fc", "mean.ratio" (in Table 4.2) etc, could distinguish between beating and gunshot spatters at BT distance 30 cm based on feature engineering data with a high accuracy, while it is difficult to find such a feature or a set of features at BT distance 120cm. In addition, when the BT distance d is unknown, it is challenging to classify the spatters at different BT distances correctly.



(a) Misclassification rate.

(b) Out-of-bag error.

Figure 4.15: Monte Carlo study with 1000 runs for the proposed classifier at different BT distance ranges using(a) misclassification rate and (b) Out-of-bag error.

Table 4.3: Performance of the model

BT distance (cm)	number of spatters	oob error (Average)	MSC rate (Average)	accuracy (Average)
$d \leq 30$cm	18	1.19%	1.04%	98.81%
$d \leq 60$ cm	40	6.80%	7.08%	93.20%
$d \leq 120$ cm	58	14.04%	14.00%	85.96%

Fig 4.16 shows the 6 most important features selected by random forest and Fig 4.17 shows oob error based on a one time simulation for spatters at BT distances $d \leq 120$ cm. The vertical axis in Fig 4.16 are features defined in Table 4.2 vs. the mean decrease in Gini index James et al. (2013). The features are sorted in decreasing order of importance. Though the rank of important features varied with different training sets which were randomly selected, most top features were common in 1000 simulations, e.g. Diameter.std, fd and fc.

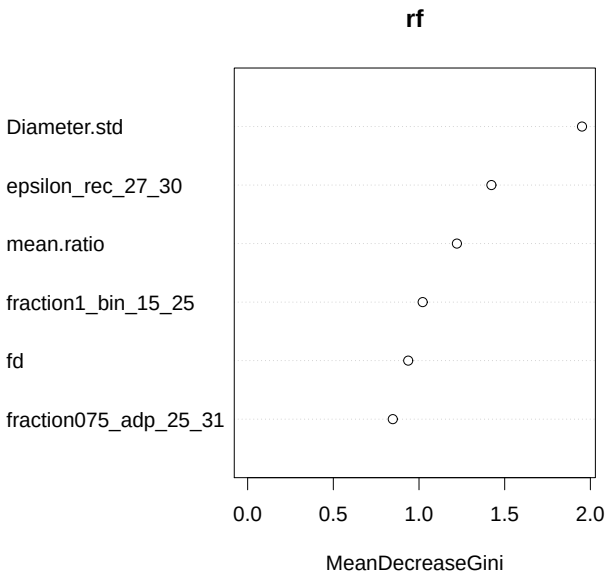


Figure 4.16: Top 6 important features

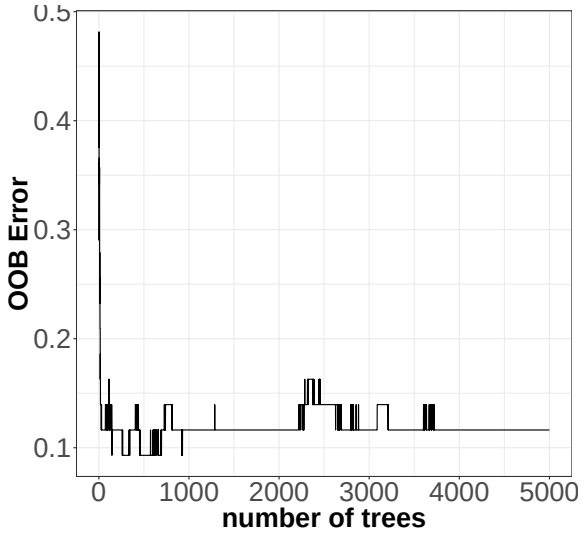


Figure 4.17: Oob error for all trees up to the ith in one simulation.

4.5.4 Further discussions

4.5.4.1 Include feature engineering data

If we use 94 spatters (i.e. including the 36 feature engineering spatters) in the Monte Carlo study to split for the training(75%) and test(25%) set, similar error rates were obtained in Table 4.4 after 1000 simulations.

Table 4.4: Performance of the model including the feature engineering data

BT distance (cm)	number of spatters	oob error (Average)	MSC rate (Average)	accuracy (Average)
$\leq$ 30cm	30	0.28%	0.14%	99.72%
$\leq$ 60 cm	64	8.00%	8.40%	92.00%
$\leq$ 120 cm	94	13.90%	13.60%	86.10%

Based on the Monte Carlo study with 1000 runs, the proposed model achieves 99.72% accuracy at BT distance $\leq$ 30cm.

4.5.4.2 Performance on gunshot spatters involving muzzle gases

Recall in Section 4.2, the data set Attinger et al. (2018b) contains 10 additional gunshot backspatters in which muzzle gases are present. We did not use those spatters in previous sections for the sake of simplicity. This section provides the performance of the proposed model on gunshot backspatters with muzzle gases. Among the 10 gunshot backspatters involving muzzle gases, the number of spatters at each BT distance are given in Table 4.5:

Table 4.5: Number of gunshot backspatters involving muzzle gases

BT distance(cm) / number of spatters	30	60	120
	2	4	4

We use random forest on 94 spatters with 58 features listed in Table 4.2 to train the model and predict the class for 10 muzzle gasses gunshot spatters, the misclassification rate is 0% for spatters at BT distance 30 and 60 cm, and 100% at BT distance 120cm. This indicates that the proposed features and model also works on gunshot backspatters with muzzle gasses at short and medium BT distances(30, 60cm), however it does not work for gunshot backspatters with muzzle gasses at BT distance $d = 120$cm. This is because those spatters muzzle gases have enough time to deflect the smaller drops which are signs of gunshot spatters

4.6 Conclusion

In this paper, we propose a machine learning method to classify between blood spatters generated by gunshot and beating, generated at various horizontal distances between blood source and stained target surface. Of the 94 spatters 36 spatters are used for feature engineering and the rest are used for training and testing. We proposed 58 features and several binning methods to segment images to construct local features. Random forest is used to

fit to the training set and to predict the label on the test set. Based on 1000 replications of random splitting training and test data, the proposed method achieves 98.81% accuracy in classifying generation mechanisms at BT distance no larger than 30cm, 93.20% accuracy to classify spatters at BT distance no larger than 60cm and 85.96% accuracy on spatters at BT distances no larger than 120cm. The main contributions of this paper are: (1) Propose a formal quantitative and physics-based method to construct new features, especially the local features. (2)Propose a novel and quantitative method to classify bloodstain spatters by combining the digital image processing method and a machine learning method.

Finally, we used the model to predict the process of gunshot backspatters involving muzzle gases. It turns out the model achieves the accuracy of 100% at BT distance 30 and 60cm and the accuracy of 0% at BT distance 120cm.

4.7 References

Arthur, R. M., Cockerton, S. L., de Bruin, K. G., and Taylor, M. C. (2015). A novel, element-based approach for the objective classification of bloodstain patterns. *Forensic science international*, 257:220–228.

Arthur, R. M., Humburg, P. J., Hoogenboom, J., Baiker, M., Taylor, M. C., and de Bruin, K. G. (2017). An image-processing methodology for extracting bloodstain pattern features. *Forensic science international*, 277:122–132.

Attinger, D., Liu, Y., Bybee, T., and De Brabanter, K. (2018a). A data set of bloodstain patterns for teaching and research in bloodstain pattern analysis: Impact beating spatters. *Data in Brief*, 18:648–654.

Attinger, D., Liu, Y., Faflak, R., Rao, Y., A. Struttman, B., De Brabanter, K., M. Comiskey, P., and L. Yarin, A. (2018b). A data set of bloodstain patterns for teaching and research in bloodstain pattern analysis:gunshot backspatters. *accepted by Data in Brief*.

Bevel, T. and Gardner, R. M. (2008). *Bloodstain pattern analysis with an introduction to crime scene reconstruction*. CRC press.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Camana, F. (2013). Determining the area of convergence in bloodstain pattern analysis: a probabilistic approach. *Forensic science international*, 231(1-3):131–136.

Comiskey, P., Yarin, A., and Attinger, D. (2017a). comiskey2017high. *Forensic science international*, 276:134–141.

Comiskey, P., Yarin, A., and Attinger, D. (2017b). Hydrodynamics of back spatter by blunt bullet gunshot with a link to bloodstain pattern analysis. *Physical Review Fluids*, 2(7):073906.

Comiskey, P., Yarin, A., and Attinger, D. (2018). Theoretical and experimental investigation of forward spatter of blood from a gunshot. *Physical Review Fluids*, 3(6):063901.

Cutler, A., Cutler, D. R., and Stevens, J. R. (2012). Random forests. In *Ensemble machine learning*, pages 157–175. Springer.

de Bruin, K. G., Stoel, R. D., and Limborgh, J. (2011). Improving the point of origin determination in bloodstain pattern analysis. *Journal of forensic sciences*, 56(6):1476–1482.

De Chazal, P., Flynn, J., and Reilly, R. B. (2005). Automated processing of shoeprint images based on the fourier transform for use in forensic science. *IEEE transactions on pattern analysis and machine intelligence*, 27(3):341–350.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, volume 66. CRC Press.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA:.

Hakim, N. and Liscio, E. (2015). Calculating point of origin of blood spatter using laser scanning technology. *Journal of forensic sciences*, 60(2):409–417.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

James, S. H., Kish, P. E., and Sutton, T. P. (2005). *Principles of bloodstain pattern analysis: theory and practice*. CRC Press.

Karger, B., Rand, S., Fracasso, T., and Pfeiffer, H. (2008). Bloodstain pattern analysis-casework experience. *Forensic science international*, 181(1-3):15–20.

Kirk, P. L. (1955). Affidavit regarding state of ohio vs. samual h. sheppard.

Kirk, P. L. and Kirk, P. L. (1953). *Crime investigation: physical evidence and the police laboratory*. Interscience New York.

Knock, C. and Davison, M. (2007). Predicting the position of the source of blood stains for angled impacts. *Journal of forensic sciences*, 52(5):1044–1049.

Laber, T., Kish, P., Taylor, M., Owens, G., Osborne, N., and Curran, J. (2014). Reliability assessment of current methods in bloodstain pattern analysis. *National Institute of Justice, US Department of Justice*.

Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomforest. *R news*, 2(3):18–22.

MacDonell, H. L. and Bialousz, L. F. (1971). *Flight characteristics and stain patterns of human blood*. National Institute of Law Enforcement and Criminal Justice.

MATLAB (2018). *R2018b*. The MathWorks Inc., Natick, Massachusetts.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076.

Pearson, K. (1895). Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, pages 832–837.

Siu, S., Pender, J., Springer, F., Tulleners, F., and Ristenpart, W. (2017). Quantitative differentiation of bloodstain patterns resulting from gunshot and blunt force impacts. *Journal of forensic sciences*, 62(5):1166–1179.

Taylor, M. C., Laber, T. L., Kish, P. E., Owens, G., and Osborne, N. K. (2016). The reliability of pattern classification in bloodstain pattern analysis, part 1: Bloodstain patterns on rigid non-absorbent surfaces. *Journal of forensic sciences*, 61(4):922–927.

Varney, C. R. and Gittes, F. (2011). Locating the source of projectile fluid droplets. *American Journal of Physics*, 79(8):838–842.

CHAPTER 5. CONCLUSION AND FUTURE WORK

In this paper, we have introduced two theoretical work and one application of the nonparametric curve estimation. We first show how to construct the simultaneous confidence intervals for local polynomial regression curve based on the volume-of-tube formula in Chapter 2, then propose a nonparametric method to estimate the derivatives based on the difference quotient in Chapter 3. Last we apply the nonparametric curve estimation techniques to construct features used in the blood pattern analysis in Chapter 4. There are tremendous demands on further research and exploration to improve the work in this thesis and we discuss them chapter by chapter.

In Chapter 2, we derive the asymptotic properties of the uniform confidence intervals for local polynomial regression based on the volume-of-tube formula. The asymptotic width of the proposed confidence interval does not change from p even to its consecutive p odd degree in the interior region. In addition, the asymptotic order of the width is slower for p even in the boundary region than in the interior while it remains the same in the boundary region for p odd. Based on the simulation study, we confirm that the proposed uniform confidence intervals attain the nominal coverage and we compare the proposed uniform confidence interval with other three types confidence intervals. In general these intervals tend to be wider than the ones based on the volume-of-tube formula.

In the context of penalized splines, Wiesenfarth et al. (2012) extended the work of Krivobokova et al. (2010) to additive models with heteroscedastic errors. We believe that extensions of our work to handle heteroscedastic and correlated errors offer interesting directions for future research, including the extension to multivariate data under the previous settings.

In Chapter 3, we propose a nonparametric method for derivative estimation in random design without estimating the regression curves. We discuss the asymptotic properties for both noisy derivative estimators and smoothed derivative estimators, and provide a tuning method to choose the parameter(s) for the first and second order noisy derivatives. In the simulation study, we compare the proposed model with local polynomial regression and smoothing splines. It turns out that the proposed model has similar performance compared to other two models.

The tuning parameters are selected in two steps in order to avoid the estimation of the correlation function. However, from the asymptotic properties of the smoothed derivative estimation, finding an efficient way to tune h and k simultaneously would greatly benefit the rate of convergence of the proposed methodology. De Brabanter et al. (2018) show that for bandwidth selection in the regression problem, minimizing asymptotic squared error is asymptotically equivalent to minimizing the residual sum squared with a kernel satisfying $K(0) = 0$. In this way, the estimation of the correlation structure can be avoid. Simultaneously tuning h and k will greatly benefit the rate of convergence of the proposed estimator. A possible lead could be the use of semi-variograms.

One drawback of the proposed framework is that the proposed first and second order derivative estimators require the estimation of the density f and distribution F. It is clear that estimating those two unknown quantities introduces extra errors in the derivative estimators. Another important research area is to find a method to adapt the proposed framework directly for arbitrary distributions without transformation.

In Chapter 4, we propose an automatic classification of bloodstain spatters generated by either gunshot or blunt impact using machine learning methods. We construct 58 features from the images and use random forest to fit the training set and to predict the label on the test set. The BT distance is a hidden variable and we did not use this information in the classification. Based on the simulation study, the proposed method achieves more than 90% accuracy in classification for bloodstain spatters generated at the short and medium

(30 cm, 60 cm) BT distance. It is difficult to classify the spatters generated at BT distance 120 cm due to the fact that small stains cannot travel further and stain size is an important feature to distinguish spatter caused by gunshot and blunt impact. We believe there is room to improve the model accuracy for spatters generated at 120 cm and beyond.

5.1 References

De Brabanter, K., Cao, F., Gijbels, I., and Opsomer, J. (2018). Local polynomial regression with correlated errors in random design and unknown correlation structure. *Biometrika*, 105(3):681–690.

Krivobokova, T., Kneib, T., and Claeskens, G. (2010). Simultaneous confidence bands for penalized spline estimators. *Journal of the American Statistical Association*, 105(490):852–863.

Wiesenfarth, M., Krivobokova, T., Klasen, S., and Sperlich, S. (2012). Direct simultaneous inference in additive models and its application to model undernutrition. *Journal of the American Statistical Association*, 107(500):1286–1296.